DOCUMENT RESUME

ED 222 560 TM 820 715

AUTHOR Romberg, Thomas A.; And Others

TITLE Construct Validity of a Set of Mathematical Superitems. A report on the NIE/ECS Item Development Project.

INSTITUTION Education Commission of the States, Denver, Colo. National Assessment of Educational Progress.; Wisconsin Center for Education Research, Madison.
SPONS AGENCY National Inst. of Education (ED), Washington, DC.
PUB DATE Jan 82
NOTE 98p.

EDRS PRICE MF01/PC04 Plus Postage.
DESCRIPTORS Cognitive Processes; Elementary Secondary Education; Item Analysis; *Mathematics Achievement; Mathematics Curriculum; Mathematics Instruction; Models; *Quantitative Tests; *Test Construction; Test Interpretation; Test Items; *Test Validity
IDENTIFIERS National Assessment of Educational Progress; *NIE ECS NAEP Item Development Project

ABSTRACT

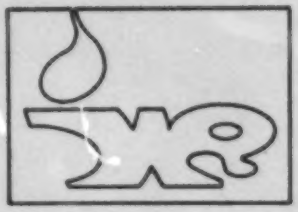
Procedural documentation is presented for administering, scoring and analyzing data gathered to examine the construct validity of a set of superitems developed to assess student levels of mathematical reasoning ability. Each superitem includes a mathematical situation and a structured set of questions about that situation. The questions were based on Collis and Biggs' recently-developed taxonomy on the structure of the observed learning outcomes (SOLO). The assumption underlying this report is that the response patterns of students to the superitems would be interpretable. To judge interpretability, three primary questions about the response patterns were raised. For each question the data strongly support the validity of the construct. The response patterns of the majority of items matched the assumed latent hierarchical and cumulative cognitive dimension. The question profiles for student clusters were interpreted in terms of developmental base stages and the spiral notions of equilibration. These findings give support to the validity of the sequence of SOLO levels. The SOLO interpretation of responses appears useful for educators and researchers in describing the level of reasoning on school related tasks. Item analyses tables are appended. (Author/PN)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED222560

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

- This document has been reproduced as received from the person or organization originating it.
- Minor changes have been made to improve reproduction quality.
- Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.



A REPORT ON THE NIE/ECS ITEM DEVELOPMENT PROJECT

Construct Validity of a Set of Mathematical Superitems

by Thomas A. Romberg, Murad E. Jurdak, Kevin F. Collis, and Anne E. Buchanan

TM 820 715

January 1982

Wisconsin Center for Education Research
an institute for the study of diversity in schooling

A report on the NIE/ECS Item Development Project

CONSTRUCT VALIDITY OF A SET OF MATHEMATICAL SUPERITEMS

by Thomas A. Romberg, Murad E. Jurdak,

Kevin F. Collis, and Anne E. Buchanan

Thomas A. Romberg
Principal Investigator

Wisconsin Center for Education Research
The University of Wisconsin
Madison, Wisconsin

January 1982

This material is based upon work supported by the National Institute of Education and the Education Commission of the States under Contract No. 02-81-20321 with the Wisconsin Center for Education Research. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of ECS, NIE, or the Department of Education.

TABLE OF CONTENTS

	Page
List of Tables	v
List of Figures	ix
Abstract	xi
Introduction	1
Construct Validity	1
The SOLO Taxonomy	2
Rules of Correspondence Between the Theory and Data	7
Other Questions	15
Test Administration	16
Construction of Items	16
Description of the Tests	17
Sample	22
Data Collection	24
Follow-up Validity Check	25
Analysis Plan	25
Results	29
Conclusions	79
References	81
Appendix A: Item Analysis Tables	83

List of Tables

Table		Page
1	Base Stage of Cognitive Development and Response Description	5
2	Guttman True-Type Response Patterns for a SOLO Superitem	10
3	Order of Superitems by Form for 17-Year-Olds. . .	19
4	Order of Superitems by Form for 9-, 11-, and 13-Year-Olds	20
5	Order of Ten Superitems Randomly Selected for Booklet 2	21
6	Assignment of Test Questions to Test Forms by Reasoning Level	23
7	Coefficient of Reproducibility (r), Probability of Misclassification (p), and χ^2 for Each Superitem by Form--17-Year-Olds	31
8	Percent Correct by SOLO Level for the Six Superitems Which Have Significant Probability of Misclassification--17-Year-Olds.	32
9	Coefficient of Reproducibility (r), Probability of Misclassifcation (p), and χ^2 for Each Superitem by Form--13-Year-Olds	34
10	Percent Correct by SOLO Level for the Eight Superitems Which Have Significant Probability of Misclassification--13-Year-Olds.	35
11	Coefficient of Reproducibility (r), Probability of Misclassification (p), and χ^2 for Each Superitem by Form--11-Year-Olds	36
12	Percent Correct by SOLO Level for the Twelve Superitems Which Have Significant Probability of Misclassification--11-Year-Olds.	38
13	Coefficient of Reproducibility (r), Probability of Misclassification (p), and χ^2 for Each Superitem by Form--9-Year-Olds.	39

List of Tables (continued)

Table		Page
14	Percent Correct by SOLO Level for the Ten Superitems Which Have Significant Probability of Misclassification--9-Year-Olds.	40
15	Summary of Questionable and Unsatisfactory Superitems	42
16	Scale Means for 17-Year-Olds on U, M, R, and E for Each Form.	43
17	Scale Means for 13-Year-Olds on U, M, and R for Each Form.	43
18	Scale Means for 11-Year-Olds on U, M, and R for Each Form.	44
19	Scale Means for 9-Year-Olds on U, M, and R for Each Form.	44
20	p values for Superitems C3 and D2 by Level of Question for Each Age Group.	47
21	Percent Correct on Each Level for Cluster Groups, 17-Year-Olds, Forms 1 to 5	48
22	Percent Correct on Each Level for Cluster Groups, 17-Year-Olds, Sample for All Forms	50
23	Percent Correct on Each Level for Cluster Groups, 13-Year-Olds, Forms 1 to 5	51
24	Percent Correct on Each Level for Cluster Groups, for 13-Year-Olds, Sample from All Forms.	52
25	Percent Correct on Each Level for Cluster Groups, 11-Year-Olds, Forms 1 to 5	53
26	Percent Correct on Each Level for Cluster Groups, 11-Year-Olds, Sample from All Forms.	55
27	Percent Correct on Each Level for Cluster Groups, 9-Year-Olds, Forms 1 to 5.	56
28	Percent Correct on Each Level for Cluster Groups, 9-Year-Olds, Sample from All Forms	57

List of Tables (continued)

Table		Page
29	ANOVA for Scale Means on Booklet 2--17-Year-Olds. .	59
30	Assignment of 17-Year-Old Students in Advanced Courses to Test Forms	59
31	ANOVA for Scale Mean for Independent Groups across Booklets--17-Year-Olds.	61
32	ANOVA for Differences in Scale Means for Dependent Groups across Booklets--17-Year-Olds.	61
33	ANOVA for Scale Means on Booklet 2--13-Year-Olds. .	62
34	ANOVA for Scale Means for Independent Groups across Booklets--13-Year-Olds	62
35	ANOVA for Differences in Scale Means for Dependent Groups across Booklets--13-Year-Olds.	63
36	ANOVA for Scale Means on Booklet 2--11-Year-Olds. .	64
37	ANOVA for Scale Means for Independent Groups across Booklets--11-Year-Olds.	64
38	ANOVA for Differences in Scale Means for Dependent Groups across Booklets--11-Year-Olds.	65
39	ANOVA for Scale Means on Booklet 2--9-Year-Olds . .	66
40	ANOVA for Scale Means for Independent Groups across Booklets--9-Year-Olds	66
41	ANOVA for Differences in Scale Means for Dependent Groups across Booklets--9-Year-Olds	67
42	Cureton's KR-20 Reliability Coefficient for Tests Made Up of Superitems for the Forms Given to 17-Year-Olds.	69
43	Cureton's KR-20 Reliability Coefficient for Tests Made Up of Superitems for the Forms Given to 13-, 11-, and 9-Year-Olds.	70
44	Readability Indices for Each Superitem (Stem and Four Questions) by Form	71

List of Tables (continued)

Table		Page
45	Readability Indices for Each Superitem (Stem and Unistructural Level Question) by Form.	73
46	Item Means and Coefficients of Reproducibility for Items Selected for the Validity Check.	75
47	Percent Correct on Seven Booklet 1 Superitems and Two Interview Superitems for 17-Year-Olds.	77
48	Percent Correct on Seven Booklet 1 Superitems and Two Interview Superitems for 13-Year-Olds.	77
49	Percent Correct on Seven Booklet 1 Superitems and Two Interview Superitems for 11-Year-Olds.	78
50	Percent Correct on Seven Booklet 1 Superitems and Two Interview Superitems for 9-Year-Olds	78

List of Figures

Figure		Page
1	Example of a Superitem Written to Reflect the SOLO Taxonomy	8
2	Basic Data Matrix for Student Responses to a Set of Superitems Based on the SOLO Taxonomy.	9
3	Four Profiles of p Values for Transition to Neighboring Cognitive Levels.	13
4	Profiles of p Values on the U, M, and R scales for 13-, 11-, and 9-Year-Olds on Booklet 1 Form 2 . . .	45

Abstract

The purpose of this report is to document the procedures followed in administering, scoring, and analyzing data gathered to examine the construct validity of a set of superitems developed to assess student levels of reasoning ability.

Each superitem includes a mathematical situation and a structured set of questions about that situation. The questions were based on a recent taxonomy of learned outcomes. The assumption underlying this report was that the response patterns of students to the superitems would be interpretable. To judge interpretability, three primary questions about the response patterns were raised. For each question the data strongly support the validity of the construct.

Thus, we conclude that we were able to construct a valid and useful set of superitems.

Introduction

The purpose of this document is to report the steps that were followed to examine the construct validity of a set of mathematical superitems. A "superitem" is a set of test questions based on a common situation or stem (Cureton, 1965). In this project, a pool of mathematical problem-solving situations and a set of items for each situation which were designed to provide information about students' qualitatively different levels of reasoning ability were developed. The structure for the questions within the superitems were based on Collis and Biggs' (1979) SOLO taxonomy used to classify the structure of observed learning outcomes. The items were prepared to be administered to students of 9, 11, 13, and 17 years of age. An earlier report describes how the items were developed (Romberg, Collis, Donovan, Buchanan, & Romberg, 1982). This report examines the construct validity of the superitems and the utility of the procedure for large scale assessments.

The project was funded by the Education Commission of the States (with funds supplied by the National Institute of Education). Ostensibly the resulting items would be useful in future National Assessment of Education Progress (NAEP) studies in mathematics.

Construct Validity

The notion of construct validity implies that the scores on a test can be meaningfully interpreted in terms of related concepts from a psychological theory. The theoretical concepts are called "constructs," and the process of validating such an interpretation is called "construct validation" (Cronbach, 1960).

Torgerson (1958) has argued that

> science can be thought of as consisting of theory on the one hand and data (empirical evidence) on the other. The interplay between the two makes science a going concern. The theoretical side consists of constructs and their relations to one another. The empirical side consists of the basic observable data. Connecting the two are rules of correspondence which serve the purpose of defining or partially defining certain theoretical constructs in terms of observable data. (p. 2)

By specifying some of the rules of correspondence which connect the theory and data and examining whether or not the data satisfy the theory, one can establish whether or not the scores are interpretable.

The SOLO Taxonomy

The theory upon which this study was based was outlined by Collis and Biggs in 1979. This theory is concerned with the reasoning and judgment a student displays in using existing knowledge. The SOLO taxonomy is based upon principles of cognitive development. Most psychologists agree that, when an individual learns something, he or she interprets it in terms of his or her existing thought structures. These structures are modified and extended according to the demands placed upon the learner. By so modifying his or her thought structure, he or she constructs an increasingly complex system of rules of thinking: some rules are general, applying to a variety of situations, while others are specific to the subject matter learned. While this process is continuous from infancy to adulthood, certain general stages of cognitive development have been distinguished. The five stages of Collis and Biggs used to describe the stages in children's judgment and reasoning ability were adapted from Piaget by Collis (1975):

1. pre-operational stage (5 to 6 years)
2. early concrete operational stage (7 to 9 years)
3. middle concrete operational stage (10 to 12 years)
4. concrete generalization stage (13 to 15 years)
5. formal operational stage (16 years onwards)

The development from pre-operational to formal operational runs the gamut from the judgment of a situation made on the basis of superficial appearances to one based on highly abstract principles.

The difficulty in directly applying these notions to school learning was that

> Piaget observed his stages of cognitive development under rather "ideal" conditions involving individual testing on quite clear-cut tasks involving general logical concepts, and so his stages tend to outline the upper limit of intellectual functioning. When we take performance in school subjects that require specific knowledge, we get rather a different picture (Collis & Biggs, 1979, p. 13).

They argued that the response a student makes to a typical school task is more complex. In fact only under ideal conditions could the level of response to such tasks be equivalent to the student's stage of cognitive development. Most often the level of response is much lower for a variety of reasons such as lack of knowledge of prerequisites and lack of interest in the subject. Furthermore, they argued, like Case (1979), that when confronted with new or unfamiliar content, one's initial reasoning about that content will be several stages lower than would be demonstrated with familiar content.

Based on this reasoning, they proposed a way of describing responses to typical school tasks, those tasks for which a student is given a

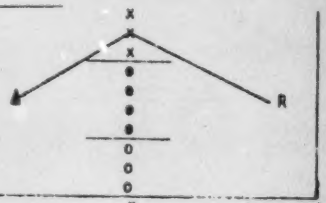
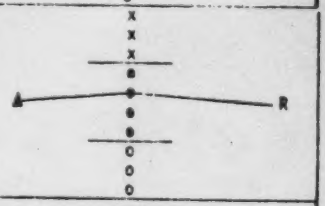
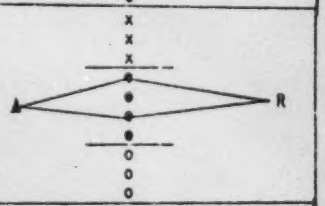
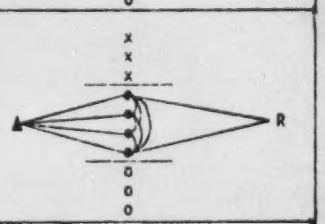
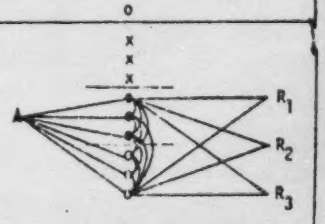
specific, finite set of information (a story, a problem in mathematics, a set of data describing and examining a concept or principle, a poem) and, on the basis of prior learning, is to answer comprehensive questions to show that the data, their interrelationships, and their possible relevance to other concepts were understood. Collis and Biggs chose to call this analysis the structure of the observed learning outcome (SOLO) to emphasize that the responses a student makes to school content reflect more than level of cognitive development.

The relationship between cognitive development and SOLO, and the general characteristics of the latter, are outlined in Table 1. At the extreme left is given the developmental base stage, in Piagetian terms, with the minimal age level at which the stage usually occurs. Next follows the name of the corresponding SOLO level. The remaining columns provide some characteristics of each SOLO level.

Capacity refers to the availability of mind-space, or more technically, working memory, that the different levels of SOLO require. Functional working memory capacity increases with age, as does the space required for higher level responses. Relating operation refers to the way in which the cue, and the aspects of the response, relate together. Consistency and closure refers to two opposing needs felt by the learner: one is the need to come to a conclusion of some kind (to close); the other is to make consistent conclusions so that there is not contradiction either between the conclusion and the data, or between different possible conclusions. The greater the felt need to come to a quick decision, the less information will be utilized, so that the probability that the outcome will be inconsistent with the original cue, the data, or the outcome is increased.

Table 1

Base Stage of Cognitive Development and Response Description

Developmental Base Stage with Minimal Age	SOLO Description	1 Capacity	2 Relating Operation	3 Consistency and Closure	4 Possible Response Structure Cue / Response
Pre-operational (4 - 6 years)	Pre-structural	Minimal: cue and response confused	Denial, tautology, transduction. Bound to specifics	No felt need for consistency. Closes without even seeing the problem	
Early Concrete (7 - 9 years)	Uni-structural	Low: cue + one relevant datum	Can "generalize" only in terms of one aspect	No felt need for consistency, thus, closes too quickly: jumps to conclusions on one aspect, and so can be very inconsistent	
Middle Concrete (10 - 12 years)	Multi-structural	Medium: cue + isolated relevant data	Can "generalize" only in terms of a few limited and independent aspects	Although has a feeling for consistency can be inconsistent because closes too soon on basis of isolated fixations on data, and so can come to different conclusions with same data	
Concrete Generalization (13 - 15 years)	Relational	High: cue + relevant data + inter-relations	Induction. Can generalize within given or experienced context using related aspects	No inconsistency within the given system, but since closure is unique so inconsistencies may occur when he goes outside the system	
Formal Operations (16+ years)	Extended Abstract	Maximal: Cue + relevant data + inter-relations + hypotheses	Deduction and induction. Can generalize to situations not experienced	Inconsistencies resolved. No felt need to give closed decisions - conclusions held open, or qualified to allow logically possible alternatives. (R_1, R_2 or R_3)	

KEY: Kinds of data used: x = Irrelevant or inappropriate • = Related and given in display o = Related and hypothetical, not given.

Note: From Classroom Examples of Cognitive Development Phenomena: The SOLO Taxonomy by K. F. Collis & J. B. Biggs, 1979, p. 16.

On the other hand, a high level of need for consistency ensures the utilization of more information in making a decision, so that the decision is likely to be more open. Structure is an attempt to represent these characteristics in diagrammatic form. The student may respond to the cue by using three types of data: irrelevant data (represented by x); related data which are contained in the original display (represented by ●); and data and principles which are not given but which are relevant, hypothetical and often implicit in the data (represented by 0).

For this project, we hypothesized that by using the SOLO framework one ought to be able to design items such that a series of questions based on the stem would require more and more sophisticated use of the information from the stem in order to obtain a correct result. This increase in sophistication should parallel the increasing complexity of structure noted in the SOLO categories.

Thus, as described in the report of the development of superitems (Romberg, Collis, Donovan, Buchanan, & Romberg, 1982), the construction of the items consisted of two parts, writing the stem and constructing questions to reflect the SOLO levels. So that a correct response to each question would be indicative of an ability to respond to the information in the stem at least at the level reflected in the SOLO structure of the particular question, we used the following criteria to write questions:

Uni-structural (U)	Use of one obvious piece of information coming directly from the stem.
Multi-structural (M)	Use of two or more discrete closures directly related to separate pieces of information contained in the stem.

Relational (R)	Use of two or more closures directly related to an integrated understanding of the information in the stem.
Extended Abstract (E)	Use of an abstract general principle or hypothesis which is derived from or suggested by the information in the stem.

In each superitem, the correct achievement of question 1 would indicate an ability to respond to the problem concerned at at least the uni-structural level. Likewise success on question 2 corresponds to an ability to respond at multi-structural level, and so on.

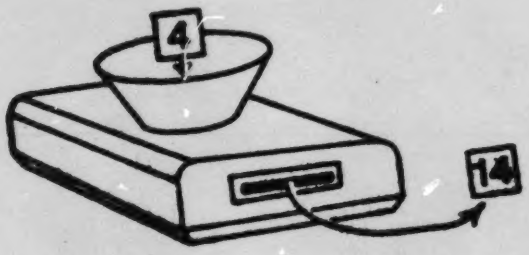
An example of items constructed in this manner is shown in Figure 1. The stem provides information and each question that follows requires the student to reason at a different level in order to produce a correct response.

Rules of Correspondence Between the Theory and Data

Superitems like the one in Figure 1, when given in group testing situations, yield a lot of data--answers and often scratch work or notes related to the answers. While much of this information could be coded, in this study only whether the answers were correct or not was coded (1 = correct, 0 = incorrect). Thus, the actual data for any student on a set of superitems are a string of 1's or 0's. Thus, if a group of students were to take a set of superitems similar to the one shown in Figure 1, a data matrix such as the one shown in Figure 2 would result.

The structure of the SOLO taxonomy assumes a latent hierarchical and cumulative cognitive dimension. Consequently, the response

This is a machine that changes numbers. It adds the number you put in three times and then adds 2 more. So, if you put in 4, it puts out 14.



U. If 14 is put out, what number was put in?

M. If we put in a 5, what number will the machine put out?

R. If we got out a 41, what number was put in?

E. If x is the number that comes out of the machine, when the number y is put in, write down a formula which will give us the value of y whatever the value of x.

Figure 1. Example of a superitem written to reflect the SOLO taxonomy.

	Superitem 1	Superitem 2	. . .	Superitem I
Student	U M R E	U M R E	. . .	U M R E
1	X X X X	X X X X	. . .	X X X X
2	X X X X	X X X X	. . .	X X X X
3	X X X X	X X X X	. . .	X X X X
.	. . . .	. . . .	. . .	. . . .
.	. . . .	. . . .	. . .	. . . .
.	. . . .	. . . .	. . .	. . . .
N	X X X X	X X X X	. . .	X X X X

X = 1 if correct, 0 if incorrect.

U = Uni-structural, M = Multi-structural, R = Relational, and E = Extended Abstract.

Figure 2. Basic data matrix for student responses to a set of superitems based on the SOLO taxonomy.

structure associated with any level of reasoning determines the response structure associated with all lower levels, in the sense that the presence of one response structure implies the presence of all lower response structures. From this we raise the following questions with respect to the pattern of responses for each item (each column in Figure 2).

Question 1. For each item is the pattern of responses a Guttman true-type response?

The five expected response patterns for each of the superitems is shown in Table 2. These five response patterns are called the Guttman true types (Guttman, 1941). Any deviation from a true type is classified as an error. A measure of the extent to which the observed response patterns belong to Guttman true types can be used to answer this first question.

Table 2

Guttman True-Type Response Patterns for a SOLO Superitem

Response Pattern	SOLO Response Level			
	U	M	R	E
1	0	0	0	0
2	1	0	0	0
3	1	1	0	0
4	1	1	1	0
5	1	1	1	1

Question 2. From their responses can the students at each age level be grouped into interpretable groups which reflect the SOLO levels?

The aggregated scores of students on superitems corresponding to the four levels of reasoning in the SOLO taxonomy provide a basis for a possible natural arrangement of subjects into homogeneous groups. If a student's responses to a set of superitems are all Guttman true-type responses, and if the student is at a particular base stage of development (see Table 1), one would expect the average response pattern across several superitems to reflect that base stage of development. It would not be expected that the response patterns would be identical for every superitem since knowledge of prerequisites, familiarity, procedural errors, and so on are also operative.

Furthermore, for a large number of students at any age level, one would expect that groups of students with similar response patterns for a set of items could be identified. It is plausible that the profiles of response patterns for the groups can be interpreted in terms of the SOLO taxonomy.

The profiles which would be interpretable are based on the notions of equilibrations which involve "formation instability combined with a progressive movement toward stability" (Langer, 1969, p. 93). Cognitive development is seen as "spiral" and, in particular, it is assumed that "to go forward it is necessary to go backward: the first step toward progress is regress" (Langer, 1969, p. 95). From a consideration of this notion, four suggested response profiles for students based on the SOLO superitems for two neighboring levels of performance are shown

in Figure 3. The first (X) and last (Y) steps show stability of performance at neighboring cognitive levels. The steps in between show the regression from X to Y (X→) and the progression from X to Y (→Y). These are seen as steps in the transition between cognitive levels.

In actuality the profiles shown in Figure 3 are ideal. The actual profiles found in this study are likely to be different for two reasons. First, because the questions at succeeding levels are more complex, there is an increase in the probability of making errors at higher levels. Thus, the p values for Y will be lower than that for X. Second, since the superitems involve different content areas and require students to read the items, either unfamiliarity with the specific content of an item or inability to read the words would depress the patterns shown in Figure 3. If particular p values on X and Y are similar but moderate, students would be reasoning at the Y level on those problems they understood. We have decided to indicate this pattern by adding the symbol "+" to its descriptors. In summary, if a student profile for the set of superitems can be grouped with other student profiles and if the groups' average profile can be judged as similar to one of the four profiles (including depressed profiles) shown in Figure 3, then interpretability in terms of a developmental base will be claimed.

Question 3. Does the superitem test format have an effect on the responses to questions at various levels?

It has been assumed that the individual questions within a superitem are not independent. In fact, it is the lack of independence that led Cureton (1965) to his discussion of such superitems. Furthermore, the

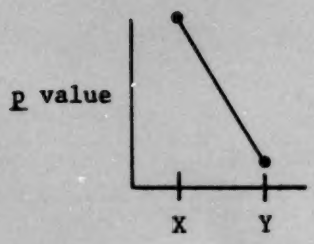
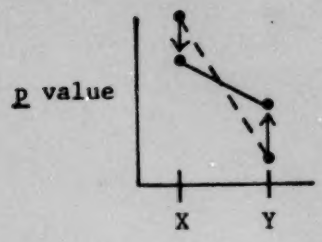
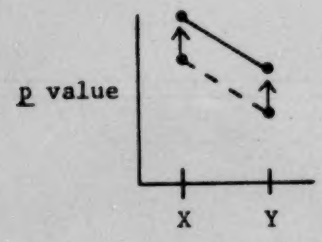
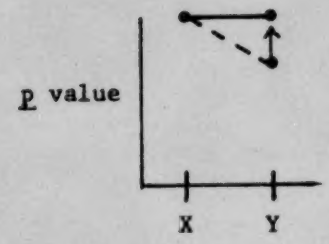
Graph	X,Y p values	Slope of line joining p value of X with p value of Y
 Step 1: Level X (X)	high X low Y	m_1: negative, steep
 Step 2: Regression to Y (X→)	decrease X increase Y	m_2: negative, shallow $\lvert m_2 \rvert < \lvert m_1 \rvert$
 Step 3: Progression to Y (→Y)	increase X increase Y	m_3: negative, shallow $\lvert m_3 \rvert = \lvert m_2 \rvert$
 Step 4: Level Y (Y)	stable X increase Y	m_4: level $\lvert m_4 \rvert < \lvert m_3 \rvert$

Figure 3. Four profiles of p values for transition to neighboring cognitive levels.

nature of the dependency needs to be determined. It is possible that asking the student a multi-structural question may focus his attention on one aspect of the problem, or asking a relational question may sugest an approach to solving an extended abstract question. Focusing the child's attention in this manner may either facilitate or debilitate the problem-solving process. We assumed that asking a lower level question would facilitate a student's response for a higher level question, but asking the higher level question first would be debilitating on the lower level question. For example, if a group of students were to take a set of multi-structural (M_1) questions, a second set of relational (R_1) questions, and a third set of two-question superitems containing both types of questions (M_2 and R_2), the group means for M questions and R questions would be $\bar{X}_{M_1} > \bar{X}_{M_2}$ and $\bar{X}_{R_2} > \bar{X}_{R_1}$.

To measure this effect, tests consisting of subsets of questions from the total set of superitems needed to be assembled. One test contained only the uni-structural question, another the multi-structural question, and so on. Other tests were also assembled, each containing two or three types of questions. From scores on these tests, it would be possible to determine if the questions had an effect on each other.

In summary, these are the three basic questions we planned to examine in this study. Answers to the first two questions are related to the construct validity of the SOLO taxonomy and the answer to the third question is about the independence of questions for the superitem format.

Other Questions

In constructing tests containing a set of superitems and administering these to a population of students, we raised these additional questions.

Question 4. What is the reliability of a test made up of superitems?

If the results of the examination of Question 3 indicate the questions do not have an effect upon one another, then the standard procedures for estimating the reliability of a test form would be appropriate. However, if the results indicate the responses to the questions are not independent, then those procedures would produce an inflated estimate of the reliability of the test.

In the event of this occurrence, then the unit for estimating the reliability will not be the individual questions but rather the superitems. The internal consistency of a test form can be estimated by KR-20 as suggested by Cureton (1965) to counter the effects of correlated errors of measurement produced by the differences among subjects in general comprehension of the item stem.

Question 5. What is the reading level of each superitem?

Since we planned to administer the same superitems to students of ages 9, 11, 13, and 17, it was reasonable to check on the reading level of the textual information in the superitems. In many cases, one could argue that inability to solve a problem might be attributed to a lack of adequate communication and comprehension rather than inability to operate on a certain reasoning level. To rule out this plausible interpretation, a readability analysis was undertaken.

Question 6. What is the relationship of a student's pattern of responses on a group-administered superitem test with his/her pattern on similar items given in an interview situation?

Under the assumption that data gathered in individual interview situations are more valid than data gathered in less costly group testing situations, we decided to see if the patterns of responses differed in the two situations. In fact, we assumed that the interview scores would be slightly higher because reading or procedural errors can be corrected and that the patterns of responses would reflect the same underlying base stage of development.

In summary, we believe that the construct validity of a set of superitems can be established by answering these six questions which relate the SOLO theory to observable data.

Test Administration

Construction of Items

The development and preliminary validation of the set of superitems is fully documented elsewhere (Romberg et al., 1982). The final set of 39 items represented seven content categories as follows:

Label	Content Categories	Number of Items
A	Numbers and Numeration	8
B	Variables and Relationships	8
C	Size, Shape, and Position	6
D	Measurement	6
E	Statistics and Probability	7
F	Unfamiliar	4
		39

One B item which had been particularly enjoyable for students at all ages and had discriminated well among levels of reasoning was chosen for the sample item (see Figure 1). Three of the most difficult items were judged appropriate only for 17-year-olds; three of the items easiest for 17-year-olds replaced them for the 9-, 11-, and 13-year-olds. Thus, there were 35 items available for assignment to test batteries for the 17-year-olds and 35 for 9-, 11-, and 13-year-olds.

Description of the Tests

Separate group-administered tests were prepared for the 17-year-olds and for the 9-, 11-, and 13-year-olds. Separate tests were necessary because in addition to the difference in items noted above, the tests for the 17-year-olds included the questions for all four levels of reasoning (U, M, R, E), whereas the tests for the three lower age levels did not include the extended abstract question. The two tests were further organized in two booklets to accommodate most conveniently the two formats in which the items would be administered (Booklet 1 to gather data to examine Questions 1 and 2 and Booklet 2 to answer Question 3). The separate booklets also were designed to discourage students from referring to previous work on an item and to allow efficient scoring and data processing.

Booklet 1 contained items in the basic superitem format. Five test forms of seven items each were created for each of the two age groups (17-year-olds and 9-, 11-, and 13-year-olds) by randomly assigning items, with the restriction that each content category except F be represented at least once but not more than twice per form. The assignment was

adjusted so that items in the same content category were not contiguous within each form. The assignment of items is outlined in Tables 3 and 4. Since the intent of the testing was to validate items, rather than to measure individual achievement, time limits were not set in the usual sense. However, based on the trial administration of the items, it was suggested that 40 minutes total (about 5-6 minutes per item) was sufficient time for most students. Booklet 1 also included directions, sample items, and a mathematics attitude questionnaire adapted from Nimier, Galmiche, and Mandrille (1980).

In Booklet 2 the items contained the stem and a question at a single level of reasoning or the stem and two questions in one of the three possible pairwise combinations of three levels of reasoning. For 17-year-olds the items contained the stem and level(s) M, R, E, MR, ME, or RE. Level U was not included in Booklet 2 for this age group although it was in Booklet 1. Using levels U, M, and R, similar items were constructed for 9-, 11-, and 13-year-olds. Thus, there were six forms of Booklet 2 for each age group, with forms containing only levels M and/or R common to both groups. Each form contained the same 10 items, randomly selected with stratification according to content categories from the 32 superitems in Booklet 1 common to both age groups. The 10 items, in the order of presentation, are listed in Table 5. Because each item had one or two fewer parts than in Booklet 1, about 4 minutes per item was suggested or 40 minutes total. Booklet 2 also contained a short, timed verbal scale adapted from the Similar Words Test (Romberg & Wilson, 1969) and the NAEP student questionnaires for the appropriate age level.

Table 3

Order of Superitems by Form for 17-Year-Olds

Question Number	Form				
	S1	S2	S3	S4	S5
1	C6	C5	F3[a]	C3	B7
2	B2	D6	E7	B6	D1
3	E8	A3	C2	D3	E3[a]
4	F1[a]	D2	D4	B4	F2
5	D5	E6	B5	E5	C1
6	B8	B3	C4	A8	A1
7	A4	E1	A6	F4	E4

[a]Item not included in tests for 9-, 11-, and 13-year-olds.

Table 4

Order of Superitems by Form for 9-, 11-, and 13-Year-Olds

Question Number	Form				
	UMR1	UMR2	UMR3	UMR4	UMR5
1	A3	B4	C2	F4	C5
2	B3	F2	B5	A6	A2[a]
3	D2	E7	D4	B7	E5
4	E6	B8	E4	A7[a]	B2
5	C6	A1	A5[a]	E1	A4
6	D1	C4	E8	D3	D5
7	A8	D6	B6	C3	C1

[a]Item not included in tests for 17-year-olds.

Table 5

Order of Ten Superitems Randomly Selected for Booklet 2[a]

Question Number	Item
1	D3
2	B7
3	B2
4	A3
5	E1
6	C1
7	D2
8	A6
9	C5
10	E6

[a]Item order is the same in all forms (U, M, R, E, UM, UR, MR, ME, RE).

The level of reasoning tested in all forms of the two booklets is outlined in Table 6. The five forms of Booklet 1 and six forms of Booklet 2 for each age group were systematically paired to ensure approximately equal numbers of all possible pairs. Individual student packets containing the two booklets were then randomly packed for distribution.

Copies of the directions for students in Booklets 1 and 2, the sample item, and the accompanying administrator's manuals appear in Romberg et al. (1982)

Sample

A central Wisconsin school district serving a community of 32,000 and the surrounding rural area agreed to provide a sample of approximately 300 students in each age group for the administration of the batteries. The school district is comprised of 2 high schools, a middle school, and 13 elementary schools. The entire grade 12 population of 310 students at one high school was tested; an additional sample of 56 students was selected from the second high school to ensure a sufficient final number of 17-year-olds. Because the middle school administrators viewed the testing as a desirable learning experience for all students, the entire grade 8 population, primarily 13-year-olds, of 562 students was tested. Of the 13 elementary schools, 8 were randomly selected to participate, providing a sample of 405 grade 6 students, 11 years old, and 323 grade 4 students, 9 years old.

The school district administrators were extremely cooperative in making arrangements for the testing, particularly in establishing a

Table 6

Assignment of Test Questions to Test Forms by Reasoning Level

Form	SOLO Response Level			
	U	M	R	E
		Booklet 1		
S1	X	X	X	X
S2	X	X	X	X
S3	X	X	X	X
S4	X	X	X	X
S5	X	X	X	X
UMR1	X	X	X	
UMR2	X	X	X	
UMR3	X	X	X	
UMR4	X	X	X	
UMR5	X	X	X	
		Booklet 2		
U	X			
M		X		
R			X	
E				X
UM	X	X		
UR	X		X	
MR		X	X	
ME		X		X
RE			X	X

positive attitude toward the testing among students and parents. This was especially important for the grade 12 students who were not required to participate. A letter publicizing the testing and encouraging full support was sent by direct mail to every parent. After reductions due to absences, underage/overage students, and a few cases of unusable data, the final sample sizes were:

Age	Number
17	303
13	490
11	370
9	308

Data Collection

The tests were administered during the week of September 14-18, 1981. Test packets containing the two booklets were randomly distributed to students. At the high schools, R&D Center staff members assisted by school staff administered both booklets during the first three class periods of one school day with the students assembled in several large group areas; there were two one-hour sittings with a short break between sittings. The additional students from the second high school were tested during two mathematics class periods by the classroom teacher. The mathematics teachers in the middle school administered the tests during math class times on three consecutive days; both questionnaires and the verbal scale were given the first day followed by the actual tests on the second and third days. At the elementary schools, the two booklets were administered in two one-hour sittings on consecutive days by classroom teachers or by the building principal. The building principals

for all participating schools completed the NAEP Principal's Questionnaires providing background information on students such as socioeonomic status and mathematics course experiences.

Follow-up Validity Check

The validity of the responses generated in the group-administered test setting was examined about six weeks after the initial administration by means of individual clinical interviews conducted with 12 students at each age level. Each student was administered two superitems.

In summary, data from 300-500 students at each of four ages, 17, 13, 11, and 9 years were collected via two booklets containing a sample of the constructed items. These data along with follow-up interview data from a small sample were used to answer the six questions being addressed in this study.

Analysis Plan

In this section, the technical procedures that were followed to examine the three primary questions and three additional questions are presented.

Question 1. For each item is the pattern of responses a Guttman true-type response?

Three indices were used to examine whether the responses of students at each age level for each superitem were true Guttman types. First, a coefficient of reproducibility was calculated in the following way:

$$\text{coefficient of reproducibility (r)} = 1 - \frac{\text{total no. of errors}}{\text{total no. of responses}}$$

Any response pattern which is not a true Guttman type (see Table 2) is considered an error. Thus, if there are no patterns which are considered errors, the coefficient of reproducibility is 1 and the scale is a perfect

Guttman scale. If all response patterns are errors, then the coefficient is obviously zero.

In addition, Proctor (1970) formulated a probablistic representation of the observed data in order to base the acceptability of Guttman method on statistical criteria of goodness of fit rather than judgment and experience. Based on maximum likelihood procedures, a misclassification parameter (p) is calculated. This is based on the predicted distribution of response types. Then, the goodness of fit of the observed distribution of types to the predicted one is investigated by chi-square techniques. The overall chi-square value is found by summing the chi-square values for all pattern differences between predicted and observed frequencies.

A scalogram analysis using Proctor's modification (SAS, 1979) was done for each superitem separately for the four age-group populations. The scale had four points (U, M, R, and E) for the 17-year-old population and three points (U, M, and R) for each of the 9-, 11-, and 13-year-old age groups.

In summary, for each superitem administered in this study, three indices are reported for each age level--a coefficient of reproducibility (r), a probability of misclassification (p), and an overall chi-square for differences between observed and predicted frequencies for the patterns of responses.

Question 2. From their responses can the students at each age level be grouped into interpretable groups which reflect the SOLO levels?

The maximum hierarchical clustering method (Johnson, 1970) was used to partition the students on each form and across forms into homogeneous

groups based on score vectors whose four components were the aggregated scores on the four taxonomic levels of reasoning: uni-structural (U), multi-structural (M), relational (R), and extended abstract (E). Before this analysis was carried out, items which failed to reflect a Guttman scale, as a result of answering Question 1, were omitted. Different possible number of cluster groups were considered and then profiles of means for each cluster group on each level of question were contrasted. These profiles and contrasts were then examined to see if they were interpretable (see Figure 3). Clusters were first found for each form and then a sample across forms for each age group.

Question 3. Does the superitem test format have an effect on the responses to questions at various levels?

To examine the effect of asking several questions based on the same item stem, three different ANOVAs were performed. From Booklet 2 responses only, an ANOVA was carried out in which the data for each type of question from each form were contrasted. In this manner, the effect of form could be tested. Again we assumed that answering a lower level question would facilitate answering a higher level question correctly, but being asked to answer a higher level question would debilitate answering a lower level question correctly.

Second, since Booklet 2 was given after Booklet 1 for all students. the effect of sequence could also be examined via two additional analyses of variance. The second ANOVA compared means for each reasoning level for independent groups of students on Booklets 1 and 2. For example, this analysis compared the mean on M questions for students on Booklet 1 who did not take any M questions in Booklet 2, with the mean of students

who did have M questions in Booklet 2. The third ANOVA compared the difference scores for students who had the same level of questions in both Booklets 1 and 2. For example, the difference in mean M scores between both booklets were found for students who took M, UM, and MR forms in Booklet 2. Again we assumed means for the higher level questions in Booklet 2 would be higher than they were for Booklet 1.

Question 4. What is the reliability of a test made up of superitems?

For each form in Booklet 1, Cureton's (1965) adaptation KR_{20} was found:

$$\frac{K}{K-1}\left[\frac{\sum_{1}^{k}\sigma_i^2}{\sigma_t^2}\right]$$

where

K = number of superitems

σ_i^2 = score-variance on each superitem. The score on a superitem is the sum of scores.

σ_t^2 = score-variance on all superitems.

Question 5. What is the reading level of each superitem?

To examine the readability of the texts of the superitems, all mathematics terms were deleted before the text was entered into a textual analysis computer program (STAR, 1978) which provides four readability indices. The Flesch Index (Flesch, 1948) is a predicted score based on average word length, in syllables, and average sentence length, in words, with a range from 0, practically unreadable to 100, easy for any literate

person. The Dale Index (Dale, 1948) is a predicted score based on average sentence length and number of unfamiliar words, those not on the Dale List of 3000 words. The FOG Index (Gunning, 1952) is based on average sentence length and number of high caliber words, words of three or more syllables. The Fry Index (Fry, 1967) is based on average number of sentences and the average number of syllables. The FOG and Fry indices are grade-level equivalent, and the Dale Index includes correction tables which give the grade equivalents.

For the 17-year-old population, these indices were based on the total superitem (stem and four questions) administered; for the 17-year-olds to answer E questions, some new information in those questions needed to be understood. For the other populations, these indices were based only on the stem and U question, which contained the basic information which needed to be read and understood.

Question 6: What is the relationship of a student's pattern of responses on a group-administered superitem test with his/her pattern on similar items given in an interview situation?

Since a very small sample of students were individually interviewed, only descriptive information has been provided about the relationship between a student's performance from the interview sittings with his/her performance in group testing.

Results

The results of data analyses are presented separately for each question. Within each question, data are presented for 17-, 13-, 11-, and 9-year-olds.

Question 1. For each superitem is the pattern of responses a Guttman true-type response?

The purpose of using Guttman scaling procedures was to examine the extent to which the latent cognitive dimension under consideration is hierarchical and cumulative. A superitem was considered to be a Guttman true-type if $r > .85$, $p \leq .15$, and χ^2 was not significant at the .05 level.

For 17-year-olds, the results of the scalogram analysis are shown in Table 7. Of the 35 superitems, 31 have a coefficient of reproducibility (r) greater than .85. This means that the errors, i.e., deviations from Guttman true types, are less than 15% of all responses for these superitems. The probabilities of misclassification (p) are not more than 15% for 30 of the 35 superitems. Consequently, there was no more than 15% response error for the items, based on the observed frequencies of nonscale types for 30 out of the 35 superitems. The goodness of fit between the actual distribution of frequencies of types and the predicted distribution based on the required probability of misclassification is given by χ^2. The χ^2 was significant at .05 with df = 10 for only six superitems, five of which had a high probability of misclassification. The six superitems for which one or more index indicated a problem are: F1, B8, E6, B3, C2, and D3.

An inspection of the six superitems and the percentage of correct responses on each of the four taxonomic levels U, M, R, E (see Table 8) reveals that for four of the superitems, the percentages are not in the

Table 7

Coefficient of Reproducibility (r), Probability of Misclassification (p), and χ^2 for Each Superitem by Form--17-Year-Olds

Superitem	r	p	χ^2
	FORM 1		
C6	.967	.031	7.2
B2	.885	.095	6.7
E8	.967	.033	2.9
F1	.877	.154[a]	23.3**
D5	.991	.008	1.5
B8	.778	.251[a]	43.1**
A4	.991	.008	6.6
	FORM 2		
C5	1.000	.001	.1
D6	.939	.005	7.4
A3	.922	.070	10.6
D2	.939	.057	11.1
E6	.784[a]	.250[a]	33.6**
B3	.810[a]	.186[a]	38.3**
E1	.948	.046	7.6
	FORM 3		
F3	.976	.020	5.4
E7	.984	.014	15.1
C2	.800[a]	.218[a]	56.8**
D4	.969	.030	2.1
B5	.907	.086	3.8
C4	.915	.083	16.2
A6	.946	.048	3.3
	FORM 4		
C3	.903	.100	11.6
B6	.982	.017	2.9
D3	.868	.140	20.4*
B4	.929	.054	14.8
E5	1.000	.001	.2
A8	.964	.032	3.0
F4	.982	.017	3.9
	FORM 5		
B7	.943	.047	6.6
D1	.983	.016	9.9
E3	.887	.100	16.6
F2	.919	.099	11.1
C1	.991	.008	2.2
A1	.943	.053	3.8
E4	.991	.008	1.6

[a]Criterion not met.

*p < .05

**p < .01

Table 8

Percent Correct by SOLO Level for the Six Superitems Which Have Significant Probability of Misclassification--17-Year Olds

Superitem	SOLO Response Level			
	U	M	R	E
F1	93.4	67.2	57.4	9.8
B8	62.3	42.6	70.5	3.3
E6	89.7	32.8	58.6	13.8
B3	96.6	65.5	46.6	63.8
C2	49.2	84.6	16.9	1.5
D3	94.7	70.2	77.2	12.3

directions predicted by the SOLO taxonomy.[1] For example, both the U and M level questions for superitem B8 were considerably more difficult than the R level item. Similarly, for three other superitems, one question is clearly out of line (R for E6, E for B3, and M for C2). Therefore, it is reasonable to argue that these four superitems have

[1]To generate the aggregated data across responses, we used the LERTAP item analysis program (Nelson, 1974). This program yields a p value for each item and several standard item characteristics such as biserial correlations with the subtest (e.g., U and M). Since interpretation of the item analysis results was not anticipated because of the nature of the test and the underlying model, these data were not examined. However, the p values and biserial correlations are reported in the appendix.

deficiencies inherent in the specific questions. These deficiencies might be attributed to improper categorization of the questions to levels, or ambiguity in the language of the items, or some other cause.

For both of the other two items (F1 and D3), the p values for the M and R questions are similar. An examination of the actual patterns indicates that the 1010 error pattern occurred considerably more often than expected (10 times for F1 and 12 for D3).

Thus, for 17-year-olds, only 4 superitems have practical problems which indicate they do not reasonably reflect the SOLO taxonomy, 2 superitems are questionable, and 29 are satisfactory.

The results of the scalogram analysis for 13-year-olds are shown in Table 9. Of the 35 superitems, 31 have a coefficient of reproducibility (r) greater than .85. The probabilities of misclassifaction (p) are not more than 15% for 32 of the 35 superitems and no significant departure from the predicted pattern was found for 28 superitems. Overall there are 8 superitems for which at least one negative indicator was found. For these 8 items, the percent correct on each of the three taxonomic levels U, M, and R is shown in Table 10.[2] For three of the items (E6, B8, and C2), one scale mean is considerably out of line with the SOLO theory and each is considered negative on each indicator. For two of the items (F2 and C4), there is no consistent Guttman pattern. Superitem D5 has a relatively easy R question which yielded 18 101 response patterns (error). The remaining superitems, C6 and A2, both have hard R questions. These yielded fewer than expected 111 patterns and hence significant χ^2's.

---

[2] Item p values and biserial correlations are reported in the appendix.

Table 9

Coefficient of Reproducibility (r), Probability of Misclassification (p), and χ^2 for Each Superitem by Form—13-Year-Olds

Superitem	r	p	χ^2
	Form 1		
A3	.987	.016	2.1
B3	.945	.066	6.3
D2	.987	.015	4.8
E6	.804[a]	.290[a]	33.1**
C6	.951	.070	7.9*
D1	.963	.038	3.7
A8	.969	.043	2.6
	Form 2		
B4	.958	.041	4.7
F2	.862	.142	20.4**
E7	.972	.034	4.3
B8	.725[a]	.709[a]	14.2**
A1	.917	.095	3.1
C4	.786[a]	.260[a]	6.9
D6	.910	.093	1.9
	Form 3		
C2[b]	.649[a]	--	--
B5	.943	.065	1.2
D4	.986	.021	.3
E4	.993	.008	.4
A5	.936	.075	2.3
E8	.915	.076	.9
B6	.950	.065	2.0
	Form 4		
F4	.957	.064	1.3
A6	.950	.060	.9
B7	.964	.042	2.8
A7	.929	.081	6.1
E1	.964	.034	1.3
D3	.873	.119	22.4**
C3	.915	.131	3.5
	Form 5		
C5	.971	.030	1.0
A2	.922	.088	9.5*
E5	.985	.016	1.7
B2	.858	.135	6.4
A4	.985	.018	2.5
D5	.992	.010	.9
C1	.964	.037	3.9

[a] Criterion not met.
[b] Item scaled very poorly so that p and χ^2 were not calculated
*p < .05
**p < .01

Table 10

Percent Correct by SOLO Level for the Eight Superitems Which Have Significant Probability of Misclassification--13-Year-Olds

Superitem	SOLO Response Level		
	U	M	R
E6	83.5	13.8	32.1
C6	84.4	88.1	.9
F2	80.4	86.6	80.4
B8	49.5	27.8	51.5
C4	60.8	59.8	32.0
C2	17.9	58.9	24.2
D3	94.7	61.1	65.3
A2	53.2	58.5	1.1

Thus for the 13-year-olds, there are 27 satisfactory superitems, 3 that are questionable, and 5 that do not reflect the SOLO taxonomy levels.

For the 11-year-olds, the results of the scalogram analysis are shown in Table 11. For the 35 superitems, 30 have a coefficient of reproducibility (r) greater than .85, 29 do not have probabilities of misclassification greater than .15 and a significant χ^2 was not found for 25 superitems. Overall there are 12 superitems for which at least one negative indicator was found. For these 12 items, the

Table 11

Coefficient of Reproducibility (r), Probability of Misclassification (p), and χ^2 for Each Superitem by Form--11-Year-Olds

Superitem	r	p	χ^2
	Form 1		
A3	.967	.045	2.5
B3	.959	.057	4.0
D2	.983	.020	14.1**
E6	.959	.049	1.2
C6	.910	.131	18.0**
D1	.926	.095	1.9
A8	.910	.153[a]	4.9
	Form 2		
B4	.981	.022	.8
F2	.759[a]	.246[a]	5.8
E7	.953	.058	3.6
B8[b]	.685[a]	--	--
A1	.907	.117	2.8
C4	.768[a]	.240[a]	10.4*
D6	.981	.016	3.1
	Form 3		
C2[b]	.699[a]	--	--
B5	.990	.008	1.4
D4	.962	.050	6.6
E4	.981	.023	1.6
A5	.915	.141	6.0
E8	.924	.072	1.5
B6	.915	.129	2.9
	Form 4		
F4	.873	.203[a]	24.0**
A6	.973	.032	.9
B7	.973	.035	9.4*
A7	.964	.039	5.5
E1	.955	.041	3.6
D3	.837[a]	.363[a]	31.1**
C3	.945	.064	1.8
	Form 5		
C5	.971	.034	1.25
A2	.906	.085	11.0*
E5	.990	.009	.9
B2	.915	.075	8.3*
A4	.990	.012	1.6
D5	.971	.041	3.1
C1	.896	.135	3.4

[a]Criterion not met.

[b]Item scaled very poorly so that p and χ^2 were not calculated.

*p < .05

**p < .01

percent correct on each of the three taxonomic levels U, M, and R is shown in Table 12.[3] For three of the items (B8, C4, and C2), one scale mean is considerably out of line with the SOLO theory and three indicators are negative. For superitems (F2 and D3), there is no consistent Guttman pattern. For the remaining superitems each has a questionable feature which produced the negative indicator. Superitems A2 and A8 have more 010 patterns than indicated. Superitems C6, F4, and A2 have hard R questions yielding no or very few 111 patterns. And Superitems D2, B7, and B2 have 1, 2, and 2 idiosyncratic 011 responses which are not expected. In the last case, we have decided that these three items actually are satisfactory.

Thus, in summary for the 11-year-olds, there are 26 satisfactory superitems, 4 questionable superitems, and 5 which do not reflect the SOLO levels.

The results for the 9-year-olds are similar. The results of the scalogram analysis for these students are shown in Table 13.[4] For 30 of the 35 superitems, r is greater than .85 and p is less than .15, and for 25 superitems χ^2 was not significant. In all only 10 items have negative indicators. For five superitems (F2, B8, C4, C2, and F4), all indicators are negative and their pattern of p values across levels is not consistent with the SOLO theory (see Table 14). Three of the items have hard R questions (p values of o). The other two items, D1 and A5, each had an idiosyncratic response of 011 which was not predicted. Thus, we decided these latter two items should be considered satisfactory.

---

[3] Item p values and biserial correlations are reported in the appendix.

[4] Item p values and biserial correlations are reported in the appendix.

Table 12

Percent Correct by SOLO Level for the Twelve Superitems Which Have Significant Probability of Misclassification--11-Year-Olds

Superitem	SOLO Response Level		
	U	M	R
D2	73.2	53.7	3.7
C6	76.8	86.6	0
A8	82.9	84.1	13.4
F2	70.8	75.0	73.6
B8	45.8	19.4	54.2
C4	40.3	25.0	34.7
C2	9.9	47.9	18.3
F4	79.7	93.2	2.7
B7	91.9	63.5	14.9
D3	94.6	43.2	35.1
A2	25.4	32.4	0
B2	15.5	15.5	8.5

Table 13

Coefficient of Reproducibility (r), Probability of Misclassification (p), and χ^2 for Each Superitem by Form--9-Year-Olds

Superitem	r	p	χ^2
	Form 1		
A3	.980	.020	3.4
B3	.951	.048	.8
D2	.971	.033	3.2
E6	1.000	.001	.1
C6	.884	.110	14.8**
D1	.990	.010	17.7**
A8	.922	.093	8.8*
	Form 2		
B4	.988	.010	1.8
F2[b]	.711[a]	--	--
E7	.933	.069	3.0
B8	.766[a]	.222[a]	28.2**
A1	.966	.037	1.2
C4[b]	.733[a]	--	--
D6	.966	.025	3.2
	Form 3		
C2	.788[a]	.165	23.2**
B5	1.000	.001	.1
D4	.933	.064	3.4
E4	1.000	.001	.09
A5	.955	.043	7.9*
E8	.900	.112	2.9
B6	.977	.023	2.9
	Form 4		
F4[b]	.759[a]	--	--
A6	.978	.021	3.9
B7	.934	.084	5.7
A7	1.000	.001	.1
E1	.989	.009	1.3
D3	.923	.085	4.2
C3	.967	.032	1.9
	Form 5		
C5	.977	.028	2.3
A2	.919	.065	9.7*
E5	1.000	.001	.1
B2	.942	.050	4.8
A4	.977	.027	2.5
D5	.988	.014	.9
C1	.965	.043	2.3

[a]Criterion not met.

[b]Item scaled very poorly so that p and χ^2 were not calculated.

*p < .05

**p < .01

Table 14

Percent Correct by SOLO Level for the Ten Superitems Which Have Significant Probability of Misclassification--9-Year-Olds

	SOLO Response Level		
Superitem	U	M	R
C6	29.9	40.0	0
D1	58.0	47.8	7.2
A8	56.5	53.6	0
F2	50.0	61.7	51.7
B8	35.0	10.0	40.0
C4	33.3	23.3	35.0
C2	0.0	26.7	6.7
A5	50.0	13.7	1.7
F4	41.0	60.7	0
A2	12.1	22.4	0

Thus, for the 9-year-olds, 27 items are considered satisfactory, 3 questionable, and 5 unsatisfactory.

In summary, for the 32 items that were administered to all four age groups, 20 were satisfactory for all ages. Furthermore, when one examines the questionable and unsatisfactory items across all ages, each appears to have a content validity problem. Only two items (B3 and F1) were questionable or unsatisfactory for just one age group (see Table 15). For both of those, the problem for 17-year-olds is only with the E question (in Superitem B3, question E is too easy and in F1 it is too hard). Thus, when one adds to the base 20 satisfactory items the three superitems only administered to 13-, 11-, and 9-year-olds we get 25 satisfactory items for those age groups. For 17-year-olds, 29 superitems were satisfactory. In general, this is strong evidence that the superitem format in which terms are constructed to fit the SOLO taxonomy forms a Guttman scale.

By contrasting the p values for each level across age levels, a consistent picture of growth can also be shown. The means for U, M, R, and E scales for each age level for each form are shown in Table 16 to 19. At each age the decrease in mean performance from U to R or E is consistent.

Furthermore, since 13-, 11-, and 9-year-olds took the same forms, a cross-sectional comparison indicates consistent growth. For example, in Figure 4, the p values for the U, M, and R scales are shown for Form 2 for three age groups. A consistent shift in performance across

Table 15

Summary of Questionable and Unsatisfactory Superitems

Superitem	Age			
	17	13	11	9
A2	NA	?	?	?
A8			?	?
B3	*			
B8	*	*	*	*
C2	*	*	*	*
C4		*	*	*
C6		?	?	?
D3	?	?	*	
E6	*	*		
F1	?	NA	NA	NA
F2		*	*	*
F4			?	*

? Questionable

* Unsatisfactory

NA Not Administered

Table 16

Scale Means for 17-Year-Olds on U, M, R, and E for Each Form

		SOLO Response Level			
Form	*n*	U	M	R	E
1	61	5.13	5.52	3.02	.64
2	58	6.28	4.45	3.01	1.38
3	65	5.91	5.43	2.29	.80
4	57	6.61	5.88	3.39	.93
5	62	6.48	5.79	3.89	.92

Table 17

Scale Means for 13-Year-Olds on U, M, and R for Each Form

		SOLO Response Level		
Form	*n*	U	M	R
1	109	6.02	4.41	1.06
2	97	5.20	4.26	2.66
3	95	5.52	4.82	1.71
4	95	5.96	4.06	1.79
5	94	5.31	4.37	1.23

Table 18

Scale Means for 11-Year-Olds on U, M, and R for Each Form

		SOLO Response Level		
Form	*n*	U	M	R
1	82	5.56	3.79	.41
2	72	4.32	2.78	2.04
3	71	4.65	3.65	1.04
4	74	5.22	3.01	.81
5	71	4.34	3.06	.41

Table 19

Scale Means for 9-Year-Olds on U, M, and R for Each Form

		SOLO Response Level		
Form	*n*	U	M	R
1	69	3.86	2.10	.12
2	60	2.95	1.70	1.45
3	60	2.92	1.50	.35
4	61	3.28	1.26	.26
5	58	3.57	2.19	.24

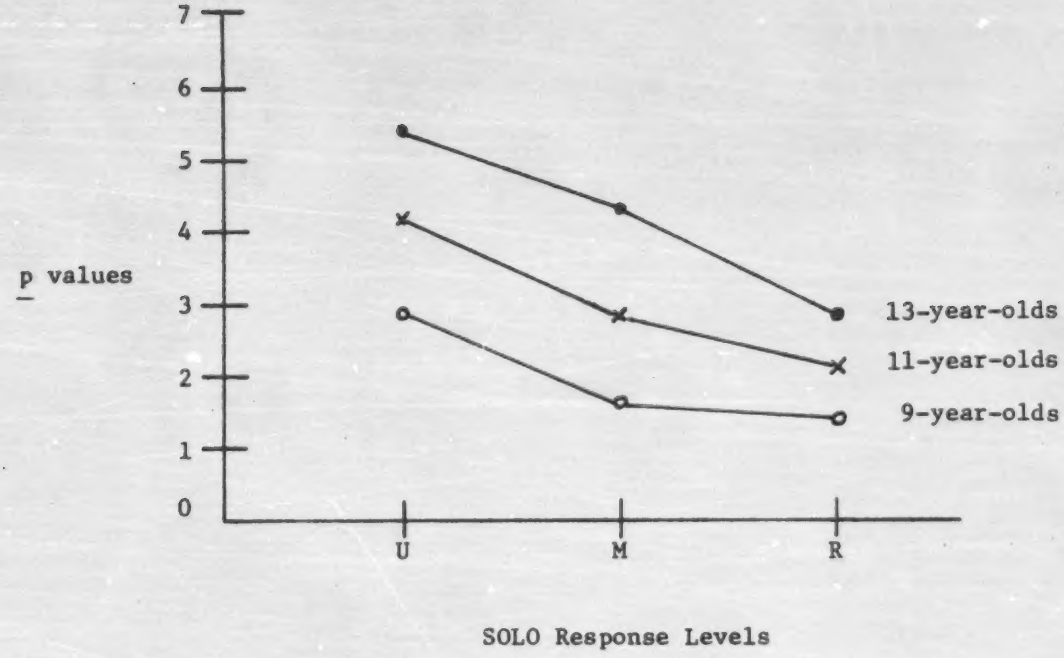


Figure 4. Profiles of p values on the U, M, and R scales for 13-, 11-, and 9-year-olds on Booklet 1 Form 2.

age levels is clear. Similarly, if one were to look at individual items, the same pattern of differences is apparent. For example, in Table 20, the p values for U, M, R, and E are shown for superitems C3 and D2 for each age group.

Although there are differences between both forms and items, the profiles of change across age levels are consistent with the notion that there are latent cognitive levels which underlie the SOLO taxonomy and that performance is cumulative and hierarchical.

Question 2. From their responses can the students at each age level be grouped into interpretable groups which reflect the SOLO levels?

For each form the mean scores for the U, M, R, and E questions across superitems were found for each student, omitting scores for the inadequate superitems. Students were then grouped via cluster analysis.

For each of the five forms of the test given to 17-year-olds, the means of each of the cluster groups are given in Table 21. In addition, the size of each cluster is given and the cluster is labeled if it was considered interpretable on the test.

The number of groups found varies from 5 to 7 depending upon the form. Twenty-seven of 28 groups over the five forms were considered interpretable. In general, the majority of students at this age are in transition between the M and R levels. A few of the students who took each form have R or higher patterns and a very few have U to M patterns.

Because of this interpretability of clusters across forms, a random sample (to be approximately 150 students) of the total population (52%)

Table 20

p values for Superitems C3 and D2

by Level of Question for Each Age Group

Superitem	Age Level	SOLO Response Level			
		U	M	R	E
C3	17	93.0	54.4	40.4	3.5
	13	88.4	30.5	14.7	--
	11	70.3	33.8	5.4	--
	9	45.9	9.8	1.6	--
D2	17	86.2	74.1	25.9	13.8
	13	89.0	74.3	14.7	--
	11	73.2	53.7	3.7	--
	9	62.3	46.4	0.0	--

Table 21

Percent Correct on Each Level for Cluster Groups, 17-Year-Olds, Forms 1 to 5

Group	n	SOLO Response Level U	M	R	E	Label
			Form 1			
1	9	96	100	62	22	R→
2	8	100	83	54	02	→R
3	14	99	95	40	17	M→
4	14	78	75	33	05	M→↓
5	11	95	86	24	04	M
6	4	87	54	13	04	→M
			Form 2			
1	9	96	91	67	44	R→
2	20	100	86	57	07	→R
3	11	78	69	24	01	M→↓
4	11	87	40	23	00	→M
5	5	48	20	00	00	U→
			Form 3			
1	11	98	97	68	41	→E
2	23	96	91	37	11	→R
3	10	81	63	45	10	M→
4	11	87	75	07	03	M
5	6	89	47	08	00	→M
			Form 4			
1	7	98	92	88	32	→E
2	17	99	94	55	21	R→
3	4	100	89	75	10	R
4	7	98	88	51	00	→R
5	10	90	64	37	05	M→
6	9	86	84	19	06	M
7	3	85	57	04	00	→M
			Form 5			
1	14	98	100	88	34	→E
2	15	92	90	67	13	R
3	16	93	83	46	05	→R
4	8	80	55	29	07	M→ ?
5	7	100	73	20	00	M

was drawn from the total population after omitting nine students whose response patterns were abnormal. The cluster analysis of the percent correct and information for this sample is shown in Table 22. The seven groups derived for the 154 students are all interpretable. Fifty-four percent of this sample are from M to R in cognitive level, 31% are above R and 16% are below M.

For 13-year-olds the derived cluster groups for each form varies from 5 to 6. The profiles for the groups for each form are shown in Table 23. Twenty-three of the 26 groups are interpretable. In general, the group profiles are around M with a few above M and approaching R, and a very few around U. Again, since the groups by form were interpretable, a random sample of 151 students was drawn (31% of the total population) across forms after six student scores were omitted.

The cluster analysis of the profiles for this sample is shown in Table 24. The largest group M comprises 50% of the sample with another 11% being M+. Fifteen percent are at level U; another 15% are above M. Finally, 6% are between U and M and 4% are below U.

For the 11-year-old population, derived cluster groups for each form varies from 5 to 7. The profiles for these groups are shown in Table 25. Twenty-eight of the 31 profiles across forms are considered interpretable. However, as one would expect because of the lower grade level, the number of "depressed" profiles for groups has increased. In general, the profiles reflect students in transition from U to M. Again, a random sample of students (154) was drawn from the population (after 7 student profiles were omitted).

Table 22

Percent Correct on Each Level for Cluster Groups
17-Year-Olds, Sample from All Forms

Group	n	SOLO Response Level				Label
		U	M	R	E	
1	25	99	94	79	35	→E
2	22	95	88	32	18	R→
3	41	97	92	58	07	R
4	30	90	76	33	01	M→
5	12	78	73	06	09	M
6	20	86	46	20	01	→M
7	4	45	20	00	00	→U

Table 23

Percent Correct on Each Level for Cluster Groups, 13-Year-Olds, Forms 1 to 5

Group	n	SOLO Response Level U	M	R	Label
		FORM 1			
1	11	.91	.74	.39	M→
2	63	.93	.82	.11	M
3	10	.65	.75	.07	M↓
4	16	.86	.42	.05	→M
5	8	.54	.31	.04	U→
		FORM 2			
1	14	.89	.98	.71	R
2	13	1.00	1.00	.34	→R
3	43	.88	.63	.18	M→
4	16	.56	.32	.16	U→
5	9	.83	.19	.00	U
		FORM 3			
1	6	.97	.89	.72	R
2	24	.97	.86	.38	→R
3	13	.87	.55	.31	M→
4	36	.95	.75	.12	M
5	15	.66	.48	.10	U→
		FORM 4			
1	12	.99	.83	.68	R
2	16	.96	.77	.38	→R
3	15	.91	.75	.16	M
4	34	.80	.51	.19	→M ?
5	12	.84	.22	.07	U
6	6	.42	.23	.02	U↓ ?
		FORM 5			
1	22	.94	.93	.38	→R
2	19	.80	.72	.23	M→
3	17	.91	.61	.08	→M
4	19	.61	.51	.09	U→
5	15	.54	.29	.04	U→↓ ?

Table 24

Percent Correct on Each Level for Cluster Groups, 13-Year-Olds, Sample from All Forms

Group	*n*	SOLO Response Level			Label
		U	M	R	
1	5	.95	.94	.76	R
2	9	.92	1.00	.43	→R
3	8	.91	.67	.50	M→
4	75	.92	.77	.15	M
5	16	.50	.48	.06	M↓
6	9	.72	.51	.13	U→
7	23	.94	.44	.02	U
8	6	.65	.14	.00	→U

Table 25

Percent Correct on Each Level for Cluster Groups, 11-Year-Olds, Forms 1 to 5

Group	n	SOLO Response Level U	M	R	Label
		Form 1			
1	3	.90	.85	.38	→R
2	6	.95	.88	.02	M
3	17	.71	.61	.08	M↓
4	11	.42	.38	.00	M↓↓ ?
5	35	.93	.57	.06	→M
6	8	.80	.25	.02	U→
7	2	.57	.07	.00	U↓ or →U ?
		Form 2			
1	2	.88	.88	.63	R
2	6	1.00	.88	.17	M
3	4	.43	.50	.00	M↓
4	9	1.00	.42	.11	→M
5	27	.75	.46	.15	U→
6	16	.59	.16	.02	→U
7	8	.16	.09	.00	P
		Form 3			
1	7	.98	.93	.43	→R
2	7	.95	.54	.29	M→
3	19	.89	.66	.07	M
4	19	.65	.47	.18	U→
5	12	.74	.28	.04	U
6	5	.23	.10	.00	P
		Form 4			
1	3	1.00	.94	.39	→R
2	2	.83	.50	.50	M→
3	18	.87	.62	.07	M
4	13	.62	.55	.10	M↓
5	24	.56	.23	.01	U→↓ ?
6	3	.90	.29	.03	U
		Form 5			
1	18	.75	.72	.14	M
2	20	.58	.46	.05	M↓
3	2	1.00	.64	.00	→M
4	20	.66	.25	.03	U→
5	10	.31	.16	.00	P

The cluster analysis of the profiles for this sample is shown in Table 26. All seven groups are interpretable. Fifty-eight percent of the population reflect a transition from U to M (U→ and →M). Another 29% have reached level M (or M↓) 11% below U, and only 9% above M.

Finally, the same procedures were followed for the 9 year old population. The number of derived cluster groups varied from 5 to 6 depending upon form. The profiles for those groups are shown in Table 27. In general, the patterns were more difficult to interpret because of the low percent correct for all R questions and most M questions and problems with "depressed" profiles. However, 23 of the 26 group profiles were considered interpretable. For students of this age, the profiles reflect patterns across the U level. A random sample of students (125 or 50%) was drawn from the population (after 7 student profiles were omitted).

The clusters formed from the profiles for this sample are shown in Table 28. All six groups are interpretable. Fifty-four percent of the students reflect a pattern around U(→U, U or U→). Twenty-eight percent are at the P level, 18% are nearing the M level.

The consistency and interpretability of the cluster profiles across the forms indicates among other things, the stable influence of cognitive levels of development in the formation of the clusters. The clusters thus formed provide support to the sequence of SOLO levels of responses.

Furthermore, the clusters strongly support the utility of the SOLO response categories over the developmental base stages in Piagetian terms. According to the taxonomy, 17-year-old students should be at formal operational level but most do not operate at the extended abstract

Table 26

Percent Correct on Each Level for Cluster Groups, 11-Year-Olds, Sample from All Forms

Group	n	SOLO Response Level			Label
		U	M	R	
1	10	.94	.91	.27	→R
2	4	.75	.46	.40	M→
3	21	.99	.68	.13	M
4	13	.42	.45	.06	M↓
5	40	.75	.62	.09	→M
6	43	.81	.36	.02	U→
7	17	.49	.18	.01	U

Table 27

Percent Correct on Each Level for Cluster Groups, 9-Year-Olds, Forms 1 to 5

Group	n	SOLO Response Level U	M	R	Label
		Form 1			
1	9	.59	.51	.08	M↓
2	20	.81	.52	.00	→M
3	4	.86	.18	.07	U
4	14	.60	.20	.00	→U or U↓?
5	13	.24	.11	.01	P→
6	8	.09	.02	.00	P
		Form 2			
1	3	.75	.58	.00	M
2	23	.40	.34	.02	M↓
3	6	.83	.17	.17	U
4	11	.56	.00	.00	→U or U↓?
5	14	.14	.00	.01	P
		Form 3			
1	5	.53	.50	.06	M↓
2	6	.89	.50	.06	→M
3	16	.75	.25	.06	U→
4	12	.42	.10	.00	→U
5	14	.13	.06	.00	P
		Form 4			
1	9	.27	.24	.03	M↓↓?
2	16	.58	.21	.04	U→
3	13	.71	.00	.01	U
4	14	.39	.02	.05	→U
5	6	.11	.00	.00	P
		Form 5			
1	5	.77	.80	.17	M→
2	5	.63	.60	.03	M↓
3	24	.55	.33	.02	U→
4	18	.38	.08	.00	→U
5	4	.14	.11	.00	P

Table 28

Percent Correct on Each Level for Cluster Groups, 9-Year-Olds, Sample from All Forms

Group	n	SOLO Response Level			Label
		U	M	R	
1	9	.52	.52	.07	M↓
2	14	.78	.57	.02	→M
3	18	.80	.36	.06	U→
4	7	.71	.00	.06	U
5	42	.47	.13	.03	→U
6	35	.16	.07	.01	P

level. For example, cluster group 1 (Table 22) has the highest performance on E questions, was judged to be at the →E stage of response, and contains only 16% of the students at this age. The majority of 17-year-olds operate around the relational level as seen from the size of clusters 1, 2, and 3 on the relational scale. This suggests that answering questions at the extended abstract level involves more than level of cognitive development.

Similar observations are obvious at each age level. No student profiles are above the hypothesized corresponding level of cognitive development. In fact, most profiles are below the base level of development for an age group. Again, this is strong support for the SOLO levels and their utility in describing responses of students.

Question 3. Does the superitem test format have an effect on the responses to questions at various levels?

To examine whether questions in the superitems are independent, one of the forms of Booklet 2 was administered to each student.

For 17-year-olds, the one-way ANOVA for differences of means on the M, R, and E scales when imbedded in different forms is shown in Table 29. Significant differences between means were found in each case. Thus, one can only conclude that the questions within the superitems are not independent. Furthermore, for the M and R scales, the means are in the predicted order. However, for the E scale, the mean for E on the RE battery, which was predicted to be the highest, is in fact the lowest.

After examining several alternatives, we attributed this discrepancy to the fact that notably few students enrolled in the advanced mathematics and science courses (calculus, pre-calculus, or physics) took form RE (see Table 30). Since one would guess these students would have more

Table 29

ANOVA for Scale Means on Booklet 2--17-Year-Olds

Scale	Test Form	*n*	Mean	Variance	F
M	M	48	.74	.042	20.8*
	MR	49	.69	.058	
	ME	51	.48	.044	
R	MR	49	.61	.057	24.5*
	R	53	.49	.058	
	RE	54	.30	.044	
E	ME	51	.36	.019	10.3*
	RE	54	.20	.024	
	E	51	.25	.052	

*p < .001

Table 30

Assignment of 17-Year-Old Students

in Advanced Courses to Test Forms

Form	Advanced *n*	Total *n*
M	10	48
R	15	53
E	32	51
MR	24	49
ME	16	51
RE	5	54

familiarity with the content the questions demand students to draw upon, their scores should be higher. The failure to have equal distribution of these students across forms could account for this lack of consistency.

The two subsequent analyses of variance to examine sequence effects for 17-year-olds are shown in Tables 31 and 32. For both the differences in means for independent groups and for dependent groups, significant F's were found on the M and E scales, but not the R scales. The Booklet 1 means are higher on the M scale and the Booklet 2 means are higher on the E scale.

For 13-year-olds, the one-way ANOVA for differences in means on the U, M, and R scales is shown in Table 33. Significant differences were found in each case and the means are in the predicted order in each case. Clearly, questions within superitems are not independent.

The two ANOVA's to examine sequence effects are shown in Tables 34 and 35. Significant differences were found for each scale for both independent and dependent groups. Furthermore, the means in both cases are in the expected order.

For 11-year-olds, the three ANOVA's are shown in Tables 36, 37, and 38. For all analyses as with 13-year-olds, significant differences were found and means are in the predicted order.

Finally, for 9-year-olds, the three ANOVA's are shown in Tables 39, 40, and 41. Significant differences in forms were found for all three scales U, M, and R and the means are in the predicted order. For sequence, significant differences were found for the U and R scales but not for the M scale. Also, for both the U and R scales, the means are in the predicted order.

Table 31

ANOVA for Scale Mean for Independent Groups across Booklets--17-Year-Olds

Scale	Test Form(s)	n	Mean	Variance	F
M	Booklet 1--All Forms	150	.75	.050	14.25*
	Booklet 2--M, MR, ME	138	.64	.061	
R	Booklet 1--All Forms	140	.46	.047	.03
	Booklet 2--R, MR, ME	148	.46	.068	
E	Booklet 1--All Forms	138	.15	.023	37.96*
	Booklet 2--E, ME, RE	150	.27	.034	

*p < .001

Table 32

ANOVA for Differences in Scale Means for Dependent Groups across Booklets--17-Year-Olds

Scale	Test Form(s)	n	Differences in Means Bklt 1-Bklt 2	Variance	F
M	Booklet 1--All Forms Booklet 2--M, MR, ME	138	.16	.047	76.73*
R	Booklet 1--All Forms Booklet 2--R, MR, RE	148	-.01	.048	.38
E	Booklet 1--All Forms Booklet 2--E, ME, RE	150	-.15	.029	113.51*

*p < .001

Table 33

ANOVA for Scale Means on Booklet 2--13-Year-Olds

Scale	Test Form	*n*	Mean	Variance	F
U	U	78	.81	.013	54.32*
	UM	80	.62	.044	
	UR	84	.52	.038	
M	UM	80	.72	.021	
	M	92	.52	.050	40.94*
	MR	75	.42	.059	
R	UR	84	.53	.025	
	MR	75	.40	.061	
	R	81	.28	.036	32.72*

*p < .001

Table 34

ANOVA for Scale Means for Independent Groups across Booklets--13-Year-Olds

Scale	Test Form(s)	*n*	Mean	Variance	F
U	Booklet 1--All Forms	244	.79	.034	62.50*
	Booklet 2--U , UM, UR	233	.65	.046	
M	Booklet 1--All Forms	235	.62	.045	12.24*
	Booklet 2--M, UM, MR	242	.55	.059	
R	Booklet 1--All Forms	243	.23	.039	77.17*
	Booklet 2--R, UR, MR	234	.40	.051	

*p < .001

Table 35

ANOVA for Differences in Scale Means for Dependent Groups across Booklets--13-Year-Olds

Scale	Test Form(s)	n	Differences in Means Bklt 1-Bklt 2	Variance	F
U	Booklet 1--All Forms Booklet 2--U, UM, UR	233	.16	.050	124.75*
M	Booklet 1--All Forms Booklet 2, M, UM, MR	242	.08	.060	25.56*
R	Booklet 1--All Forms Booklet 2--R, UR, MR	234	-.16	.050	119.60*

*$p < .001$

Table 36

ANOVA for Scale Means on Booklet 2--11-Year-Olds

Scale	Test Form	n	Mean	Variance	F
U	U	67	.71	.037	52.29*
	UM	61	.49	.038	
	UR	61	.40	.020	
M	UM	61	.58	.033	43.11*
	M	57	.32	.044	
	MR	63	.27	.042	
R	UR	61	.45	.015	59.89*
	MR	63	.24	.035	
	R	61	.15	.023	

*p < .001

Table 37

ANOVA for Scale Means for Independent Groups across Booklets--11-Year-Olds

Scale	Test Form(s)	n	Mean	Variance	F
U	Booklet 1--All Forms	176	.70	.039	52.12**
	Booklet 2--U, UM, UR	187	.54	.049	
M	Booklet 1--All Forms	187	.47	.049	9.12*
	Booklet 2--M, UM, MR	176	.39	.059	
R	Booklet 1--All Forms	183	.14	.025	55.60**
	Booklet 2--R, UR, MR	180	.28	.040	

*p < .01

**p < .001

Table 38

ANOVA for Differences in Scale Means for Dependent Groups across Booklets--11-Year-Olds

Scale	Test Form(s)	n	Differences in Means Bklt 1-Bklt 2	Variance	F
U	Booklet 1--All Forms Booklet 2--U, UM, UR	187	.14	.063	59.44*
M	Booklet 1--All Forms Booklet 2--M, UM, MR	176	.07	.057	17.68*
R	Booklet 1--All Forms Booklet 2--R, UR, MR	180	-.15	.044	92.05*

*$p < .001$

Table 39

ANOVA for Scale Means on Booklet 2--9-Year-Olds

Scale	Test Form	n	Mean	Variance	F
U	U	49	.58	.030	60.46*
	UM	52	.35	.023	
	UR	54	.28	.013	
M	UM	52	.45	.042	70.21*
	M	52	.18	.021	
	MR	51	.10	.013	
R	UR	54	.32	.022	75.88*
	MR	51	.11	.010	
	R	50	.06	.006	

*$p < .001$

Table 40

ANOVA for Scale Means for Independent Groups across Booklets--9-Year-Olds

Scale	Test Form(s)	n	Mean	Variance	F
U	Booklet 1--All Forms	152	.47	.058	8.03*
	Booklet 2--U, UM, UR	153	.40	.038	
M	Booklet 1--All Forms	151	.25	.039	.12
	Booklet 2--M, UM, MR	154	.25	.049	
R	Booklet 1--All Forms	151	.08	.017	25.00**
	Booklet 2--R, UR, MR	154	.17	.026	

*$p < .01$

**$p < .001$

Table 41

ANOVA for Differences in Scale Means for Dependent Groups across Booklets--9-Year-Olds

Scale	Test Form(s)	n	Differences in Means Bklt 1-Bklt 2	Variance	F
U	Booklet 1--All Forms Booklet 2--U, UM, UR	153	.08	.073	14.43*
M	Booklet 1--All Forms Booklet 2--M, UM, MR	154	.007	.054	.15
R	Booklet 1--All Forms Booklet 2--R, UR, MR	154	-.11	.032	60.15*

*$p < .001$

In summary for all four age groups the questions within a superitem cannot be considered independent. Furthermore, with one understandable exception, the results suggest that asking a lower level question prior to a higher level question increases performance on the latter question, and asking a higher level question descreases performance on a lower level question.

Also, a sequence effect is apparent. Responding to higher level questions goes up on the second administration of such questions, while responding to lower level questions goes down.

Question 4. What is the reliability of test made up of superitems?

The reliability estimates for each form based on Cureton's procedure are shown in Tables 42 and 43.

The estimates for forms given to the 17-years olds (see Table 42) vary from .55 to .82. These coefficients are not high but are reasonable considering that there are only 7 superitems per form and there was little variability in the U questions on all forms and the E questions on Form 1.

The estimates for the forms given the three other populations are reported separately for each age level (see Table 43). The coefficients range from .42 to .72, .48 to .74, and .35 to .71 for the 13-, 11-, and 9-year-olds, respectively. The coefficients are not high but are considered acceptable.

Question 5. What is the reading level of each superitem?

The criteria we decided to use to judge whether a superitem was too difficult for students were as follows:

For 17-year-olds (12th grade), if Flesch < 50 and one other index > 13; for 13-year-olds (8th grade), if Flesch < 70 and one other index > 9; for 11-year-olds (6th grade), if Flesch < 80 and one other > 7; and for 9-year-olds (4th grade), if Flesch < 90 and one other index > 5. The overall results of the readability analysis for superitems as administered to 17-year-olds are presented in Table 44. All superitem stems and questions were judged to be of reading difficulty appropriate to twelfth graders. Only two superitems have any index greater than 12.0 and a Flesch index in the 50's (superitem E6 on Form 2 on the FOG and superitem E5 on Form 4 on the Fry). It is interesting but not surprising that both are E items whose content is probability and statistics.

Table 42

Cureton's KR-20 Reliability Coefficient for Tests Made Up of Superitems for the Forms Given to 17-Year-Olds

Form	n	KR-20
1	61	.55
2	58	.82
3	65	.72
4	57	.72
5	62	.73

Table 43

Cureton's KR-20 Reliability Coefficient for Tests Made Up of Superitems for the Forms Given to 13-, 11-, and 9-Year-Olds

Form	13-Year-Olds		11-Year-Olds		9-Year-Olds	
	n	KR-20	*n*	KR-20	*n*	KR-20
1	109	.42	82	.60	60	.71
2	97	.65	72	.48	60	.50
3	95	.60	71	.74	60	.70
4	95	.72	74	.56	61	.35
5	94	.70	71	.59	58	.71

Table 44

Readability Indices for Each Superitem (Stem and Four Questions) by Form

Superitem	Flesch Index	Dale Index	FOG Index	Fry Index
		Form 1		
C6	77.1	7.4	7.5	7.5
B2	90.0	6.8	6.8	5.8
E8	92.9	6.6	6.0	5.2
F1	61.9	8.3	7.5	10.5
D5	95.3	6.5	4.2	3.5
B8	96.0	6.4	3.7	3.5
A4	72.6	7.2	6.9	7.6
		Form 2		
C5	78.0	7.4	9.2	7.5
D6	67.0	8.0	10.5	9.2
A3	73.0	7.7	7.6	8.5
D2	89.0	6.8	6.5	6.5
E6	56.3	8.5	12.5	11.6
B3	96.0	6.4	3.7	3.5
E1	82.4	7.2	8.1	7.3
		Form 3		
F3	71.1	7.8	9.4	7.6
E7	80.1	7.3	8.0	7.4
C2	72.9	7.7	7.3	7.6
D4	71.1	7.8	8.2	8.5
B5	103.0	6.1	2.6	2.3
C4	101.3	6.2	3.9	3.5
A6	100.2	6.2	3.8	3.3
		Form 4		
C3	90.2	6.7	5.7	5.8
B6	75.2	7.5	8.8	7.6
D3	92.0	6.7	4.7	5.2
B4	71.4	7.8	8.3	8.5
E5	53.4	8.7	9.5	13.5
A8	67.4	8.0	10.8	9.5
F4	78.3	7.4	7.5	7.4
		Form 5		
B7	87.8	6.8	6.0	6.6
D1	86.5	7.0	7.9	6.6
E3	74.9	7.6	8.6	8.2
F2	93.8	6.6	6.9	6.3
C1	83.5	7.1	6.1	6.8
A1	77.0	7.5	6.0	7.2
E4	70.3	7.8	10.2	8.7

NOTE: Item judged too difficult for 17-year-olds if Flesch <50 and one other index >13.

The overall results of the readability analyses for the stems and U questions as administered to 13-, 11-, and 9-year olds is presented in Table 45. For 13-year-olds, four superitems (A1, A5, E5, and D6) were judged to be inappropriately difficult for them and several more superitems were marginal; overall the superitems seemed appropriate for students at this age. For 11-year-olds, the readability of test items is questionable; 12 of 35 items were too difficult and several were marginal. Finally, for 9-year-olds, 24 items were judged too difficult and several were marginal.

Hence, the reading difficulty of the problem-solving superitems in their present format does not seem appropriate for 9-year-olds. They are marginally appropriate for 11-year-olds and are adequate for both 13-year-olds and 17-year-olds.

Question 6. What is the relationship of a student's pattern of responses on a group-administered superitem test with his/her pattern to similar items given in an interview situation?

The interview data were gathered on a very small sample of students, 12 at each age level, and each student was asked to respond to two superitems. The students were selected at each age level on the basis of an initial cluster analysis for two of the test forms; Form S5 for the 17-year-olds, and Form UMR5 for the 9-, 11-, and 13-year-olds. Disregarding outlying cases, four discrete clusters were identified at each age level. These cluster groups were not identical, but very similar, to those reported earlier in this report. These were from an initial analysis performed before unsatisfactory items were omitted where we also specified the number of cluster groups a priori. Three students were randomly selected from each of the four clusters at each age level.

Table 45

Readability Indices for Each Superitem (Stem and Unistructural Level Question) by Form

Superitem	Flesch Index	Dale Index	FOG Index	Fry Index
		Form 1		
A3	76.8	7.5	5.4	7.5
L3	99.6	6.3	3.1	3.3
D2	101.8	6.1	3.9	3.1
E6	70.2	7.8	11.4	8.5
C6	89.6	6.8	5.7	5.8
D1	87.7	6.9	7.7	6.6
A8	70.5	7.8	9.8	8.5
		Form 2		
B4	80.2	7.3	8.1	7.5
F2	86.5	6.9	6.9	6.4
E7	87.9	6.9	7.0	6.8
B8	70.2	7.8	6.6	8.3
A1	57.2	8.5	7.6	10.8
C4	102.8	6.1	3.9	3.2
D6	63.9	8.1	11.4	9.8
		Form 3		
C2	76.8	7.5	5.8	7.2
B5	102.0	6.1	1.3	1.5
D4	85.3	7.0	5.8	7.0
E4	77.5	7.4	7.6	7.4
A5	53.0	8.7	12.1	13.7
E8	103.5	6.0	4.6	2.9
B6	75.5	7.5	8.7	7.5
		Form 4		
F4	85.3	7.0	7.7	7.0
A6	99.6	6.3	3.1	3.3
B7	83.0	7.1	6.8	7.2
A7	92.5	6.6	7.0	6.4
E1	81.6	7.2	8.3	7.4
D3	92.6	6.6	3.8	3.9
C3	92.0	6.7	4.8	4.7
		Form 5		
C5	85.4	7.0	7.7	6.6
A2	83.7	7.1	6.8	6.6
E5	57.1	8.5	9.3	11.7
B2	93.4	6.6	6.4	6.3
A4	69.2	7.9	7.6	8.3
D5	90.7	6.7	5.1	5.2
C1	86.0	7.0	5.6	5.6

NOTE: Item judged too difficult for 13-year-olds if Flesch <70 and one other index >9, for 11-year-olds if Flesch <80 and one other index >7, for 9-year-olds if Flesch <90 and one other index >5.

The two items administered at each age level were selected according to several criteria. First, the 10 items in Booklet 2 and in the two forms (S5 and UMR5) used in the cluster analysis were eliminated from consideration since the students had attempted them previously. Second, the items had to discriminate reasonably well among levels of reasoning, insofar as this could be determined from the initial results of the item and scalogram analyses. A corollary concern here was that there be a possibility of at least a few students responding correctly at the highest reasoning level relevant for their age; for example, if no 17-year-old student had responded correctly to the extended abstract question in the group tests, the item was not considered for the interview at that age level. Finally, the two items were to be from different content areas. Statistics for the selected items are presented in Table 46.

R & D staff conducting the interviews read the stem to 9-year-olds while the 11-, 13-, and 17-year-olds read all parts of the item independently. The stem and the questions at each level of reasoning were provided on separate cards which were handed to the students one at a time. All 9-, 11-, and 13-year-olds were given both the uni-structural (U) and multi-structural (M) levels; if a student could not respond correctly or make a reasonable attempt at solving the level M question, the relational (R) level was not administered. The three 13-year-old students in the highest cluster were also given the extended abstract (E) level question, though this was not the case in the group-administered tests. The 17-year-olds were all administered the first three levels, whether or not they successfully responded to each; however, only those students who could perform

Table 46

Item Means and Coefficients of Reproducibility

for Items Selected for the Validity Check

Level	Selected Items							
	17 Yrs.		13 Yrs.		11 Yrs.		9 Yrs.	
	B5	A8	B5	A1	E7	A1	E7	A1
	Percent Correct							
U	.89	.98	.74	.92	.81	.80	.55	.68
M	.74	.96	.49	.71	.52	.69	.28	.35
R	.57	.67	.16	.36	.09	.22	.10	.06
E	.23	.19						
	Coefficient of Reproducibility							
	.907	.964	.943	.917	.953	.907	.933	.966

at the relational level were given the extended abstract questions. If the interviewer was uncertain how students determined their responses and/or to verify the level of reasoning employed, students were queried following each question.

A comparison of percent correct for groups of students on the group-administered booklet and the interview is shown in Tables 47 to 50. Overall, for the 52 comparisons, in 46 cases the interview percent correct is higher than the group administered p value. Several reasons for the differences were apparent. For U and M questions, the interviewers noted several instances where students raised questions which clarified their understanding of questions or got them to correct a procedure error. For R and E questions, prompts or answers to questions (or lack of answers) caused students to rethink the question. And for the 9-year-olds, since the questions were read to the students in the interview situation, that source of error was alleviated.

Nevertheless, the overall pattern of responses continues to strongly support the SOLO taxonomy. What it indicates is that the group testing situation adds another factor to the response level interpretation. The students in the four groups at each level are different. At ages 17-, 13-, and 9-, Group 1 performs at a higher level than the others. Groups 2 and 3 have similar performance profiles but Group 3 students asked more questions, received more prompts, etc. And Group 4 students remained low. For age 11 students, Groups 1 and 2 had similar profiles but Group 2 students needed more help and the profiles for Groups 3 and 4 were still lower. Finally, for 13-year-olds in the interview situation, we

Table 47

Percent Correct on Seven Booklet 1 Superitems and

Two Interview Superitems for 17-Year-Olds

Group	Situation	U	M	R	E
1	Booklet 1	.95	1.00	.90	.29
(n = 3)	Interview	1.00	1.00	1.00	.67
2	Booklet 1	.90	.95	.76	.19
(n = 3)	Interview	1.00	1.00	.50*	.50
3	Booklet 1	.95	.76	.38	.05
(n = 3)	Interview	1.00	1.00	.67	.50
4	Booklet 1	.90	.52	.24	.05
(n = 3)	Interview	1.00	.33*	.00*	.00*

*Interview score booklet score.

Table 48

Percent Correct on Seven Booklet 1 Superitems and

Two Interview Superitems for 13-Year-Olds

Group	Situation	U	M	R	E
1	Booklet 1	.95	.95	.24	-
(n = 3)	Interview	1.00	1.00	.50	.10
2	Booklet 1	.81	.71	.14	
(n = 3)	Interview	1.00	.83	.50	
3	Booklet 1	.86	.67	.05	
(n = 3)	Interview	1.00	.83	.33	
4	Booklet 1	.62	.48	.05	
(n = 3)	Interview	1.00	.83	.17	

Table 49

Percent Correct on Seven Booklet 1 Superitems and

Two Interview Superitems for 11-Year-Olds

Group	Situation	U	M	R
1	Booklet 1	.71	.81	.19
(n = 3)	Interview	1.00	.83	.33
2	Booklet 1	.57	.48	.14
(n = 3)	Interview	1.00	1.00	.17
3	Booklet 1	.57	.19	.10
(n = 3)	Interview	1.00	.33	.00*
4	Booklet 1	.19	.14	.00
(n = 3)	Interview	.67	.17	.00

*Interview score < booklet score.

Table 50

Percent Correct on Seven Booklet 1 Superitems and

Two Interview Superitems for 9-Year-Olds

Group	Situation	U	M	R
1	Booklet 1	.81	.76	.14
(n = 3)	Interview	1.00	1.00	.33
2	Booklet 1	.57	.62	.05
(n = 3)	Interview	.83	.83	.17
3	Booklet 1	.57	.33	.00
(n = 3)	Interview	1.00	.50	.17
4	Booklet 1	.38	.14	.00
(n = 3)	Interview	.83	.50	.00

asked the Group 1 students to try the E questions. In general, these students are not ready to answer such questions although an occasional student may be able to work problems at this level.

The group administration of this type of superitem yields a score for a student which is somewhat lower than one would get by interviewing the student. At the U and M levels, careless errors are far too common in group administration. At the R and E levels, the capability to respond correctly depends so much on the content of the particular problem that the score must be considered a lower bound for a student's level of reasoning. Thus, a student correctly answering 3 or 4 of 7 E questions should be considered able to reason at that level.

Conclusions

The purposes of the analyses related to the six questions examined in this document were to examine the construct validity of the superitems developed in this project and to estimate the utility of the testing procedure for large scale assessments.

The majority of items constructed in this project proved to be Guttman true-type items. Thus, the response patterns match the assumed latent hierarchical and cumulative cognitive dimension. Furthermore, from the question profiles for each student, clusters of students were formed and the profiles for those clusters were interpretable in terms of developmental base stages and the spiral notions of equilibration. Together these findings gave strong support to the validity of the sequence of SOLO levels.

The utility of the SOLO approach to superitem construction and interpretation of responses is also apparent. Answering content-based questions at varying levels requires more than level of cognitive development. Thus, the SOLO interpretation of responses is more useful for educators and researchers to describe level of reasoning on school-related tasks.

References

Case, R. The underlying mechanisms of intellectual development. In J. Biggs and J. Kirby (Eds.), Instructional processes and individual differences in learning. New York: Academic Press, 1979.

Collis, K.F. A study of concrete and formal operations in school mathematics: A Piagetian viewpoint. Melbourne: Australian Council for Educational Research, 1975.

Collis, K.F., & Biggs, J.B. Classroom examples of cognitive development phenomena: The SOLO taxonomy. Report prepared at conclusion of an Educational Research and Development Committee funded project, University of Tasmania, 1979.

Cronbach, L.J. Essentials of psychological testing (2nd Edition). New York: Harper & Row, 1960.

Cureton, E.E. Reliability and validity: Basic assumptions and experimental designs. Educational and Psychological Measurement, 1965, 25, 327-346.

Dale, E. A formula for predicting readability. Educational Research Bulletin, 1948, 27, 11-20.

Flesch, R. A new readability yardstick. Journal of Applied Psychology, 1948, 32, 221.

Fry, E. A readability formula that saves time. Journal of Reading, 1967-68, 11, 513.

Gunning, R. The techniques of clear writing. New York: McGraw-Hill Book Company, 1952

Guttman, L. The quantification of a class of attributes: A theory and method for scale construction. In P. Horst (Ed.), The prediction of personal adjustment. New York: Social Science Research Council, 1941, pp. 319-348.

Johnson, S.C. Hierarchical clustering schemes. Psychometrika, 1970, 35, 241-254.

Langer, Theories of development. New York: Holt, Rinehart, and Winston, 1969.

Nelson, L.R. LERTAP. University of Otago, Duneden, New Zealand, 1974.

Nimier, J., Galmiche, J., Mandrille, A. Multidimensional research on the affective attitude of the pupils towards mathematics. Paper presented at the Fourth International Conference for the Psychology of Mathematics Education. Berkeley, CA: 1980.

Proctor, C.H. A probabilistic formulation and statistical analysis of Guttman scaling. Psychometrika, 1970, 35, 73-78.

Romberg, T.A., Collis, K.F., Donovan, B.F., Buchanan, A.E., & Romberg, M.N. A report on the NIE/ECS Item Development Project: The development of mathematical problem-solving superitems. Madison: Wisconsin Center for Education Research, 1982.

Romberg, T.A., & Wilson, J.W. The development of tests. In J.W. Wilson, L.S. Cahen, and G. Begle (Eds.), NLSMA Reports Volume 7. Stanford, CA: School Mathematics Study Group, 1969.

SAS. SAS User's Guide 1979 Edition. Raleigh: SAS Institute Inc., 1979.

STAR. Textual Reading Analysis. Madison: Madison Public Schools, 1978.

Torgerson, W.S. Theory and methods of scaling. New York: John Wiley & Sons, 1958.

Appendix A

ITEM ANALYSIS TABLES

Table A1

Results of Item Analysis by Form--17-Year-Olds

Superitem	% Correct				Biserial Correlation[a]			
	U	M	R	E	U	M	R	E
				Form 1 n = 61				
C6	93.4	90.2	26.2	.0	.88	.56	.64	.00
B2	73.8	77.0	62.3	49.2	.91	.57	.82	1.04
E8	95.1	96.7	70.5	.0	.47	.47	.47	.00
F1	93.4	67.2	57.4	9.8	.88	.73	.45	.63
D5	95.1	95.1	.0	.0	.19	.19	.00	.00
B8	62.3	42.6	70.5	3.3	.99	.92	.78	.97
A4	100.0	83.6	14.8	1.6	.00	.75	.45	.87
				Form 2 n = 58				
C5	94.8	62.1	51.7	1.7	.79	.57	.48	-.11
D6	89.7	89.7	56.9	13.8	1.02	.79	.83	1.17
A3	84.5	63.8	22.4	17.2	.51	.75	.77	.92
D2	86.2	74.1	25.9	13.8	.89	.87	.79	.76
E6	89.7	32.8	58.6	13.8	.84	.80	.79	.64
B3	96.6	65.5	46.7	63.8	.96	.80	1.03	.76
E1	86.2	56.9	48.3	13.8	.81	.76	.68	1.17
				Form 3 n = 65				
F3	93.8	66.2	60.0	23.1	.28	.84	.81	1.13
E7	96.9	84.6	16.9	.0	.78	.69	.40	.00
C2	49.2	84.6	16.9	1.5	.93	.96	.68	.85
D4	98.5	95.4	13.8	10.8	.69	1.24	.81	.68
B5	89.2	73.8	56.9	23.1	.85	.83	.76	.79
C4	76.9	72.3	30.8	1.5	.74	.76	.38	.85
A6	86.2	66.2	33.8	20.0	.84	.82	.80	.89
				Form 4 n = 57				
C3	93.0	54.4	40.4	3.5	1.01	.74	.80	.54
B6	98.2	94.7	57.9	.0	.67	.62	.74	.00
D3	94.7	70.2	77.2	12.3	.63	.80	.71	.91
B4	93.0	87.7	64.9	57.9	.64	.80	.99	1.02
E5	87.7	87.7	7.0	.0	1.08	.72	.49	.00
A8	98.2	96.5	66.7	19.3	.26	.47	.74	.98
F4	96.5	96.5	24.6	.0	.35	.47	.64	.00
				Form 5 n = 62				
B7	96.8	87.1	45.2	35.5	.46	.93	.62	.74
D1	93.5	85.5	35.5	.0	.60	.69	.83	.00
E3	82.3	67.7	66.1	25.8	1.00	.95	.66	1.03
F2	88.7	93.5	79.0	6.5	.79	.49	.35	.90
C1	98.4	96.8	66.1	3.2	.50	.66	.85	1.11
A1	91.9	95.2	53.2	21.0	.72	.38	.90	.81
E4	96.8	53.2	43.5	.0	.69	.86	.69	.00

[a] The biserial correlation is for the item with the subtest (e.g., U, M, R, or E) not the form as a whole.

Table A2

Results of Item Analysis by Form--13-Year-Olds

Superitem	% Correct			Biserial Correlation[a]		
	U	M	R	U	M	R
			Form 1 n = 109			
A3	85.3	62.4	.0	.74	.54	.00
B3	88.1	55.0	22.9	.53	.71	.65
D2	89.0	74.3	14.7	.42	.54	.76
E6	83.5	13.8	32.1	.69	.63	.76
C6	84.4	88.1	.9	.75	.78	.69
D1	78.9	68.8	26.6	.66	.62	.70
A8	92.7	78.9	9.2	.72	.69	.93
			Form 2 n = 97			
B4	83.5	50.5	35.1	.77	.81	.87
F2	80.4	86.6	80.4	.52	.41	.41
E7	90.7	80.4	9.3	.60	.51	.34
B8	49.5	27.8	51.5	.55	.55	.63
A1	92.8	71.1	36.1	.38	.64	.78
C4	60.8	59.8	32.0	.77	.56	.59
D6	61.9	49.5	21.6	.83	.85	.81
			Form 3 n = 95			
C2	17.9	58.9	24.2	.66	.60	.39
B5	74.7	49.5	16.8	.79	.70	.72
D4	97.9	93.7	3.2	.14	.57	.97
E4	95.8	26.3	15.8	.94	.86	.74
A5	86.3	73.7	25.3	.78	.82	.88
E8	89.5	89.5	65.3	1.00	.43	.46
B6	89.5	92.6	20.0	.79	.77	.74
			Form 4 n = 95			
F4	92.6	88.4	9.5	.59	.33	.79
A6	89.5	50.5	21.1	.86	.79	.73
B7	90.5	73.7	23.2	.55	.67	.80
A7	66.3	58.9	11.6	.70	.82	.80
E1	73.7	43.2	33.7	.68	.78	.88
D3	94.7	61.1	65.3	.82	.78	.70
C3	88.4	30.5	14.7	.99	.48	.76
			Form 5 n = 94			
C5	96.8	46.8	44.7	.57	.49	.64
A2	53.2	58.5	1.1	.86	.67	.25
E5	67.0	67.0	.0	.80	.79	.00
B2	41.5	41.5	34.0	.88	.82	1.03
A4	90.4	53.2	8.5	.66	.81	1.02
D5	94.7	88.3	.0	.76	.79	.00
C1	87.2	81.9	35.1	.86	.63	.94

[a]The biserial correlation is for the item with the subtest (e.g., U, M, or R) not the form as a whole.

Table A3

Results of Item Analysis by Form--11-Year-Olds

Superitem	% Correct			Biserial Correlation[a]		
	U	M	R	U	M	R
			Form 1 n = 82			
A3	86.6	48.8	2.4	.93	.63	.92
B3	86.6	22.0	9.8	.61	.71	1.03
D2	73.2	53.7	3.7	.71	.84	.57
E6	73.2	8.5	6.1	.47	.57	.96
C6	76.8	86.6	.0	.77	.76	.00
D1	76.8	75.6	6.1	.60	.54	.82
A8	82.9	84.1	13.4	.86	.84	1.11
			Form 2 n = 72			
B4	77.8	18.1	5.6	.77	.64	.58
F2	70.8	75.0	73.6	.51	.57	.65
E7	81.9	52.8	9.7	.68	.76	.37
B8	45.8	19.4	54.2	.61	.51	.62
A1	80.6	69.4	22.2	.52	.61	.78
C4	40.3	25.0	34.7	.43	.27	.54
D6	34.7	18.1	4.2	.68	.81	.13
			Form 3 n = 71			
C2	9.9	47.9	18.3	.60	.75	.58
B5	45.1	15.5	4.2	.79	.80	.88
D4	73.2	63.4	2.8	.83	.79	.40
E4	88.7	19.7	8.5	.61	.60	.85
A5	83.1	60.6	5.6	.81	.87	.70
E8	84.5	77.5	49.3	.71	.70	.80
B6	80.3	80.3	15.5	.75	.66	1.05
			Form 4 n = 74			
F4	79.7	93.2	2.7	.44	.56	.94
A6	83.8	21.6	12.2	.68	.74	1.04
B7	91.9	63.5	14.9	.59	.75	.87
A7	60.8	29.7	2.7	.65	.82	.08
E1	40.5	16.2	8.1	.73	.76	.73
D3	94.6	43.2	35.1	.65	.29	.90
C3	70.3	33.8	5.4	.91	.53	.83
			Form 5 n = 71			
C5	91.5	29.6	18.3	.41	.74	1.02
A2	25.4	32.4	.0	.67	.80	.00
E5	46.5	39.4	.0	.74	.66	.00
B2	15.5	15.5	8.5	.68	.66	1.04
A4	90.1	35.2	1.4	.58	.84	.35
D5	88.7	87.3	.0	.78	.58	.00
C1	76.1	66.2	12.7	.79	.69	1.06

[a]The biserial correlation is for the item with the subtest (e.g., U, M, or R) not the form as a whole.

Table A4

Results of Item Analysis by Form--9-Year-Olds

Superitem	% Correct			Biserial Correlation[a]		
	U	M	R	U	M	R
			Form 1 n = 69			
A3	59.4	17.4	.0	.67	.70	.00
B3	49.3	4.3	4.3	.67	.23	1.30
D2	62.3	46.4	.0	.65	.75	.00
E6	71.0	.0	.0	.81	.00	.00
C6	29.0	40.6	.0	.67	.82	.00
D1	58.0	47.8	7.2	.82	.84	1.45
A8	56.5	53.6	.0	.86	.88	.00
			Form 2 n = 60			
B4	46.7	5.0	1.7	.67	.64	.22
F2	50.0	61.7	51.7	.73	.67	.82
E7	55.0	28.3	10.0	.71	.72	.58
B8	35.0	10.0	40.0	.68	.83	.77
A1	68.3	35.0	6.7	.67	.63	.53
C4	33.3	23.3	35.0	.22	.62	.60
D6	6.7	6.7	.0	.46	.85	.00
			Form 3 n = 60			
C2	.0	26.7	6.7	.00	.98	.91
B5	15.0	.0	.0	.81	.00	.00
D4	36.7	33.3	1.7	.69	.99	.51
E4	70.0	1.7	.0	.65	.75	.00
A5	50.0	13.3	1.7	.64	.57	.51
E8	60.0	51.7	23.3	.94	.85	1.08
B6	60.0	23.3	1.7	.94	.57	.51
			Form 4 n = 61			
F4	41.0	60.7	.0	.58	.84	.00
A6	47.5	4.9	.0	.50	.37	.00
B7	82.0	19.7	14.8	.55	.60	1.21
A7	29.5	1.6	.0	.62	.72	.00
E1	23.0	4.9	3.3	.26	.86	1.09
D3	59.0	24.6	6.6	.78	.64	1.00
C3	45.9	9.8	1.6	.74	.72	.58
			Form 5 n =58			
C5	89.7	20.7	12.1	.37	.90	1.42
A2	12.1	22.4	.0	.83	.95	.00
E5	25.9	22.4	.0	.96	.71	.00
B2	8.6	5.2	8.6	.34	.63	1.37
A4	72.4	34.5	1.7	.61	.76	1.79
D5	74.1	62.1	.0	.77	.71	.00
C1	74.1	51.7	1.7	.92	.74	1.14

[a]The biserial correlation is for the item with the subtest (e.g., U, M, or R) not the form as a whole.